A SPEECH ANALYSIS AND SYNTHESIS SCHEME FOR BANDWIDTH COMPRESSION

STANFORD R. WILDE

UNITED STATES
NAVAL POSTGRADUATE SCHOOL

THESIS

A SPEECH ANALYSIS AND SYNTHESIS SCHEME

FOR BANDWIDTH COMPRESSION

* * * * *

Stanford R. Wilde

12ND P2238 (1-59)

A SPEECH ANALYSIS AND SYNTHESIS SCHEME FOR BANDWIDTH COMPRESSION

by

Stanford R. Wilde
Lieutenant, United States Navy

Submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

IN

ENGINEERING ELECTRONICS

United States Naval Postgraduate School
Monterey, California

1 9 5 9

A SPEECH ANALYSIS AND SYNTHESIS SCHEME FOR BANDWIDTH COMPRESSION

by

Stanford R. Wilde

This work is accepted as fulfilling
the thesis requirements for the degree of

MASTER OF SCIENCE
IN
ENGINEERING ELECTRONICS
from the
United States Naval Postgraduate School

ABSTRACT

Speech processing schemes which result in a reduced transmission bandwidth for voice communications have been the subject of intensive investigation in recent years. This paper describes a new speech analysis-synthesis scheme for bandwidth reduction. The speech analyzer develops seven analogue control signals from the speech signal. These control signals require a total bandwidth of approximately 140 cps for transmission to the synthesizer which utilizes the control signals to continuously synthesize artificial speech.

The writer wishes to express his appreciation for the assistance and encouragement given him by Professor Mitchell L. Cotton of the U. S. Naval Postgraduate School and for the original suggestion and assistance given him by William C. Dersch of the International Business Machines Corporation Research Laboratory at San Jose, California.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

1. Naval Tactical Communications System.

The exchange of tactical information within operational units of the Naval Establishment has for many years been centered around voice communications. But just as the manner in which warfare is conducted changes, so must change the means by which operational information is exchanged. There exists one basic criteria by which the means of exchange for information of this type must ultimately be judged. This criteria is: Does the means operate as an enhancement or as a constraint on the current manner in which warfare is conducted. It is of paramount importance that the means of communication in no way restricts naval tactics or the full use of current naval weaponry. The tremendous scope of naval warfare, the extreme destruction and speeds involved in current weapons and their manner of delivery, and their requirements of versatility, flexibility, and mobility on naval tactics create requirements on operational communications which are of the most stringent and severe character.

The inadequacy of today's voice communication system in meeting the demands for a tactical information exchange media has been obvious for some time. Voice communication information exchange rates are completely insufficient to cope with the problems of modern day air defense. The extreme bandwidth requirements of voice communication has long ago led to an unfulfilled demand for tactical communication channels. The acute shortage of frequency spectra caused by the use of extremely wide bandwidth channels is a problem which must be solved. The advent of the various Tactical Data Systems has been a direct consequence of this voice communication inadequacy. And with the impact of the Tactical Data System upon the naval communication scene a re-evaluation of voice communications

is inevitable.

Consider the scope of operations in which the Navy must perform. The Navy is involved in air, sea, underwater, and assault landing operations. The Navy is concerned with guided missile submarine operations, hunter-killer antisubmarine operations, fast carrier operations with air attack capabilities, assault landings across defended beaches using the concept of air envelopment, air defense against both guided missiles and manned aircraft, and a myriad of other operations. The Naval Establishment does and must have some capability in every type of warfare known to man. The Navy must be able to conduct all of these operations anywhere in the world, not from fixed, but from highly mobile bases, and in an extremely short amount of time.

Dispersion of naval forces became a necessity with the advent of thermonuclear devices. High-speed aircraft and missiles have made the reaction time both for offensive and defensive operations critically short.

What then are the demands today upon a naval tactical communication system? The system must handle tremendous amounts of varied information. It must handle this information quickly and reliably over far greater distances than ever before. It must do all this while operating under a very serious constraint. That constraint is the limited electromagnetic frequency spectrum available to naval forces.

The Tactical Data Systems are a great step toward the fulfillment of these demands. But no data system complex can handle more situations than those for which it is built. Data system complexes are built to handle a given number of situations. If an enemy so conducts his military operations such that they are not one of the given number of situations, then other means must be available for information exchange on

his operations.

Consider a data system complex which might be created to handle a combined Navy-Marine assault across a defended beach. No complex could be created to handle the information connected with every eventuality, every variation that the operation might take. True, a complex can be created to handle a great deal of the information connected with an assault landing. But it is impossible to categorize or even know every bit of information that might have an exchange requirement. And if every variation is not known, then the system cannot be designed to handle it. This philosophy is equally applicable to HUK operations, ASW, air defense, etc.

It appears apparent that every data system complex must have associated with it some means for handling what might loosely be called the unexpected variations of warfare. This flexibility in the Tactical Communication System is deemed to be extremely critical. No potential enemy can be considered so unprofessional as not to take immediate advantage of any lack of flexibility in our communication system. It is believed that voice communication as a mode of information exchange provides the most flexible communications capability.

Is then a tactical naval communication system to be burdened not only with the prodigious number of voice nets now required, but also a number of data system complexes? This writer considers the answer to be in essence, yes.

In brief, the observations made thus far are:

1. Voice communication, as known in the Naval Establishment today, is no longer adequate to serve as the primary means of tactical information exchange.

2. Data system complexes are replacing voice communications as the primary media for exchange.

3. Design limitations on data system complexes and the vital requirement of communication flexibility require that there be associated with data system complexes a communication mode possessing great flexibility.

4. Voice communications possesses great flexibility.

5. There exists an extremely critical shortage of available frequency spectra.

6. Bandwidth occupancy of additional frequency space by data system complexes make the communication picture completely untenable.

From these observations, it may be concluded that tactical communications will be carried out by data system complexes which will be supplemented by voice communications, and that the transmission voice signals must be accomplished using very much narrower bandwidths than are now occupied.

2. Criteria For Voice Communications Systems.

This paper is an investigation of a newly conceived speech processing technique and the development of circuitry to achieve the required speech processing. The particular line taken by the investigation and the goals aimed at are based upon a set of criteria which are considered to be applicable to a military voice communication system. The role played by the voice communication system is considered as a supplement to data system complexes and an integral part of an over tactical communication system.

First, the required bandwidth for the voice channel must be as small as possible, subject to other considerations. The intelligibility of the system must be firmly based upon individual word recognition by the human receiver at the output end. High intelligibility scores on connected text are not considered adequate. For, in connected text, the mind has the unique ability to fill in isolated, unrecognized words based on the line of thought of the text. A major part of naval voice traffic consists of prowords, individual code words, and in general unconnected text where the absolute recognition of words is essential. Word recognition is a basic must and as such acts as a constraint on the level of bandwidth compression achievable. Bandwidth compression involving compression in the time domain possesses undesirable attributes. Systems of this type involve time delays. Although the delays involved are usually small, it is felt that in an era of Mach two or three aircraft, a voice system which has no time delay between the input to a voice channel and the output, is a more preferable system. It was felt that the investigation should thus proceed into "no delay" systems.

Speech processing adds additional components to a conventional voice transmission system. In one direction speech must proceed through a speech analysis component, through a transmitting device, a receiving device, and a speech synthesizer. The actual speech analysis and synthesis devices may be identical for all military services. These devices should also be compatible with any system of transmission or modulation scheme. Thus, the speech processing units should work equally well whether the voiced information is sent via SSB, AM, FM, with any modulation scheme, delta modulation, frequency multiplexing, or schemes of a digital nature.

Digital transmission of the speech information possesses qualities that are desirable. These qualities are increased range, improved reliability, and inherent security. Classified techniques of digital transmission offer even more attractive qualities. Digital transmission has the disadvantage of practically requiring more bandwidth than is encompassed by the sampled wave itself.

It is felt that the modulation scheme to be used in transmitting the speech information is properly the subject of a full investigation itself, and is beyond the scope of the current investigation into the processing of speech.

Weight considerations are of the utmost importance in the development of the speech processing devices. Inasmuch as these devices are additional equipment that must be carried by aircraft, etc., an extrapolation into the future state of the electronic art was made such that a sizable weight reduction over the equipment developed during this investigation should be realizable within one to two years.

The ultimate speech processing technique used in a tactical voice

communication system should provide a level of security over that which may be obtained from the modulation scheme. The particular information bearing signals at the output of a speech synthesizer should be of such a character that a compromise of the channel depends not only on a complete knowledge of the modulation scheme, but also the exact role played by the information bearing signals in the processing scheme.

A question that must be considered is whether the speech processing scheme should be of such a character as to permit individual voice recognition. The degree of bandwidth compression obtainable in speech processing is a direct function of this speaker recognition level.

Several factors must be considered. It is a known fact that it is possible to determine individual ship location and movement from the recognition of CW operators by their particular traits. Inasmuch as the number of voice communicators is reasonably small, speaker recognition provides an easy means for ship recognition. A degree of security is thus provided by having a system in which all voices sound alike.

Contrariwise, with non-recognition, it is impossible to tell an enemy voice from a friendly voice. This is not felt to be a strong counter-argument for even with speaker recognition, it cannot be expected that enemy voices will necessarily sound different. It is felt that authentication techniques will provide the desired security. Also, higher degrees of bandwidth compression are attainable with speaker non-recognition. A system having no speaker recognition is believed to be more desirable because of its greater advantages.

Another feature which should be included in any voice communication system is that there should be a relative silence at the terminal end of the system between words. In clipped speech systems, for instance, between

words noise generates zero crossings with the result that the output in the absence of speech is very noisy.[1]

In conclusion, the guideposts for this investigation and the criteria which are believed to form a basis for a military voice communication system are:

1. Minimum bandwidth occupancy per voice channel.
2. Word recognition.
3. No time delay from speaker to receiver.
4. Compatability of the speech processor with any mode of transmission or modulation scheme.
5. Minimum weight, and thus circuit simplicity.
6. A level of security derived from the speech processing itself.
7. Speaker non-recognition.

3. Speech Parameters and Phenomena.

A survey of the literature in the field of Speech Processing shows that much and yet little has been done. Organized scientific investigation of any magnitude in this field has been restricted in time to the last 20 years. This upsurge of research and investigation has been the direct result of need: the need to meet the increasing demands upon communication services imposed by both civilians and the military; the need to find an economy in the means of exchange of voiced information by electronic devices. An economy is needed that is both an economy of channel bandwidth and equipment. The inefficiencies involved in the current electronic means of exchanging voiced information by transmitting a replica of the speech waveform have long been common knowledge to the communication engineer.

The field of human communication is an extremely broad one. Investigations in this field have been carried out by the psychologist, the acoustic engineer, the linquist, the phonologist, and experts in the field of communication and information theory. Common to all these lines of investigation is the vast lack of knowledge of the mechanism by which the human perceives speech. This is the basic and unsolved problem of human communication.

The human perception mechanism is a completely astounding, fascinating and little understood thing. The means by which a human is able to classify many diverse physical stimuli into the same category is an area of colossal ignorance. In the case of auditory recognition the same words spoken by a man and a woman are drastically different in their acoustic content, and yet, the listener has little difficulty in establishing they are the same word. The speech waveform for a spoken word varies from

person to person and even varies with time with a given person. The accents of various speakers, the emotional frame of the speaker all lead to an endless variety of waveforms for the same spoken word. Yet, the listener is able to correctly classify the word. The mechanism by which this auditory recognition is continuously carried out in the face of non-speechlike acoustic stimuli (wind noises, machinery noises and other environmental sounds) is little understood at the present time.

The endeavors of the various types of investigators in the field of human communication has lead to an array of hints and clues about the auditory recognition mechanism. A great number of phenomena concerning speech and its perception have been observed and reported. But all of the acquired knowledge has not led to such a level of understanding that the communication engineer may analytically design an efficient means for electronically exchanging voiced information.

The communication engineer today is attempting to solve two closely allied problems; the problem of efficiently communicating between men, and the problem of direct voice communication between man and machine. Communication between man and his machines is at present confounding some of the best scientists in the world. Progress in this area has been difficult and the results meager. Communication between men with regard to required bandwidths, reliability, etc., has progressed almost as slowly as man-machine communication with slightly better results.

The processing of speech to achieve the aforementioned economies in the electronics exchange of speech information between men is the problem of the communication engineer. These engineers utilizing the hints and clues provided by allied investigators in the field of human communication, taking cognizance of the reported phenomena and hypothesis have

achieved a certain level of success in providing devices to meet the demanded economies. One of the first of these devices developed and perhaps the most well known is the Vocoder as developed by Dudley.

The activities of the communication engineer in the area of man to man communication has been and is device stimulated. The goal has been to develop a means and a device to achieve bandwidth compression and increased reliability without a complete knowledge of human perception and communication. But really, the entire field of electronics is characterized by this type of thing. Awe inspiring progress was made by scientists who had little or no knowledge of the electron or how it performed. As a result of this viewpoint research in the speech processing field has been and is along non-analytic lines. What must be said is that we do not know enough about the field to be analytic.

Before considering the particular investigation presented in this paper, it is necessary to discuss briefly the speech production mechanism and the various hints, clues, and reported phenomena about human communication available to the researcher in the field of speech processing.

The process of speech production may be regarded as similar to that of a carrier system in which the modulation of a vocal cord tone or wide band fricative noise is effected by the movement of tongue, lips, jaws, and other parts of the articulation mechanism; and by the resonant qualities of nasal, mouth, and throat cavities.[2] The lungs supply to the larynx and its associated vocal folds the breath stream which is the driving force for the system. The current theory, as discussed by Stetson[3] is that the lungs do not supply the vocal mechanism with air at constant pressure during speech but in a pulsating manner so as to aid in syllable production. Of course, if a given speech sound is maintained for a long

period of time such as is encountered when the sound is sung then the air is supplied at a constant pressure.

The breath stream is constituted of a vast number of turbulent motions, each of minute energy, and so the driving force for the vocal cords is an acoustic spectra of uniform energy.[4] The vocal cords operating on the breath stream determine which of the two basic types of acoustic excitation is presented to the upper vocal organs for modulation. If the vocal folds remain in a fixed open position, such as does occur for fricative sounds, then the breath stream passes through the glottis (the space between the vocal folds) to be modulated by the resonant cavities of the upper vocal tract, the nasal and mouth cavities, and the teeth. The modulation of the uniform energy breath stream by these upper vocal organs results in a reinforcement of certain broad frequency regions within the spectrum of the breath stream. The sounds produced by this turbulent excitation are usually referred to as unvoiced sounds. The fricative "s" is a sound produced by turbulent excitation. Spectral analysis has shown that the areas of reinforcement are in general above 3000 cps for sounds produced in this manner.

The sound type of acoustic excitation is produced when the vocal cords or folds, as they are more correctly called, do not remain in the fixed open position but open and close periodically. The larynx contains the vocal folds and the associated muscles for controlling the mode of operation of the vocal folds. The larynx may be divided into three areas: 1. the subglottic cavity; 2. the space between the vocal folds, the glottis; and 3. the supraglottic cavity.[5] The subglottic cavity operates to concentrate the breath stream toward the glottis. The primary laryngeal tone is produced at the glottis for voiced sounds; while the supraglottic

cavity commences to form the timbre of the voice. The classical aerodynamic theory of phonation describing the mode of vocal fold vibration has in recent years become accepted. This "air puff" or "air burst" theory describes the sequence of vocal fold vibration as follows: 1. closure of the glottis; 2. accumulation of subglottic pressure; 3. explosion of the closed vocal folds and the escape of an air puff or burst through the opened glottis; 4. relaxation of the folds to the closed position; and 5. repetition of the cycle. The resulting pressure waveform at the upper end of the larynx is a rough asymmetrical sawtooth very rich in harmonics. The Fourier line spectra produced is not one of uniform energy. The lower harmonics contain most of the energy. As the harmonic number increases the associated energy decreases. The periodicity of the vocal fold burst is determined by the tension of the vocal folds.

The Fourier line spectra at the larynx during phonation is modulated by the upper vocal organs and cavities such that certain harmonics are attenuated and others are reinforced. Particular frequency regions in the spectra which are reinforced more strongly are called formants. The sounds produced by this harmonic excitation are called voiced sounds. The vowels are all voiced sounds. In general, there are three formants which occur during voiced sounds. These formants usually occur within the following frequency regions:[6]

F_1 270 to 730 cps

F_2 840 to 2230 cps

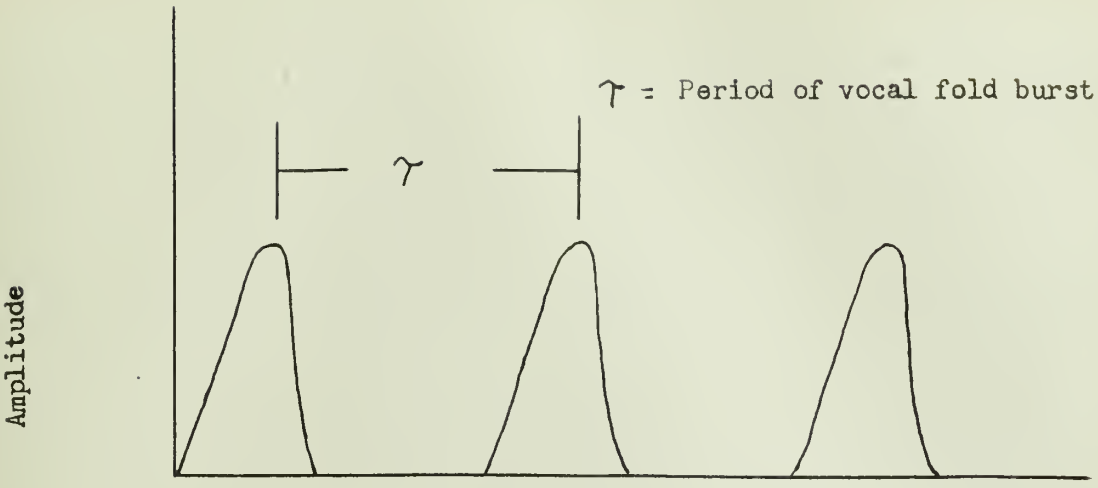
F_3 2240 to 3010 cps

The frequency corresponding to the repetition rate of the vocal fold burst is the fundamental of the Fourier series. The frequency corresponding to the pitch as heard by the listener is in most cases the fundamental

frequency of the Fourier series. In other cases the pitch frequency may be the second or third harmonic frequency.[7] Pitch phenomena and the extraction of the pitch frequency from speech by speech analyzers has plagued investigators for many years. Inasmuch as the method of pitch extraction developed and utilized in this investigation is unique, a fuller discussion of pitch will be delayed until Section 5, in which the conceptual details of the investigation conducted will be presented.

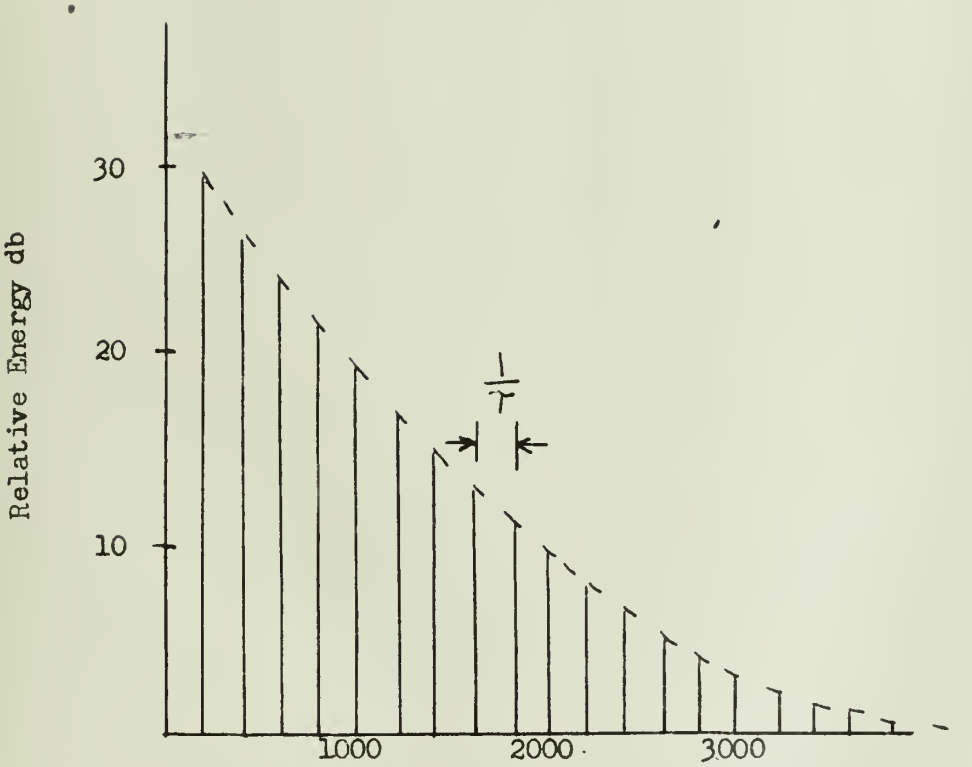
Figure 1 shows the waveform of the larynx source for voiced sounds. Figure 2 shows the approximate spectrum of the larynx source energy for a voiced sound. Figure 3 shows a typical speech waveform for voiced sounds, and Figure 4 shows the Fourier line spectra for the wave. The three formants are easily distinguished. It should be noted that the larynx harmonics may or may not lie exactly at the same frequency as the peaks of the formants.[4]

The starting point for all electronic communication systems whose function is to provide a means for the exchange of spoken information is the acoustic pressure wave generated at the lips of a speaker. Communication engineers working in voice communications have conducted analysis of speech in both the time and frequency domains. The results of these investigations has shown that while the analyzed speech of an individual speaker is directly correlatable to the operation of his vocal organs, the correlation between the observed phenomena in the time and frequency domains for different speakers is far from satisfactory. Sixty persons may say the vowel "a" and the associated pitch of the sound may be different for all. An important part of the vowel sounds is the position of the formants. The formants for a given sound shift up and down in the frequency domain depending on whether the speaker is male or female. Unfortunately, the formants do not keep the same relative positions as



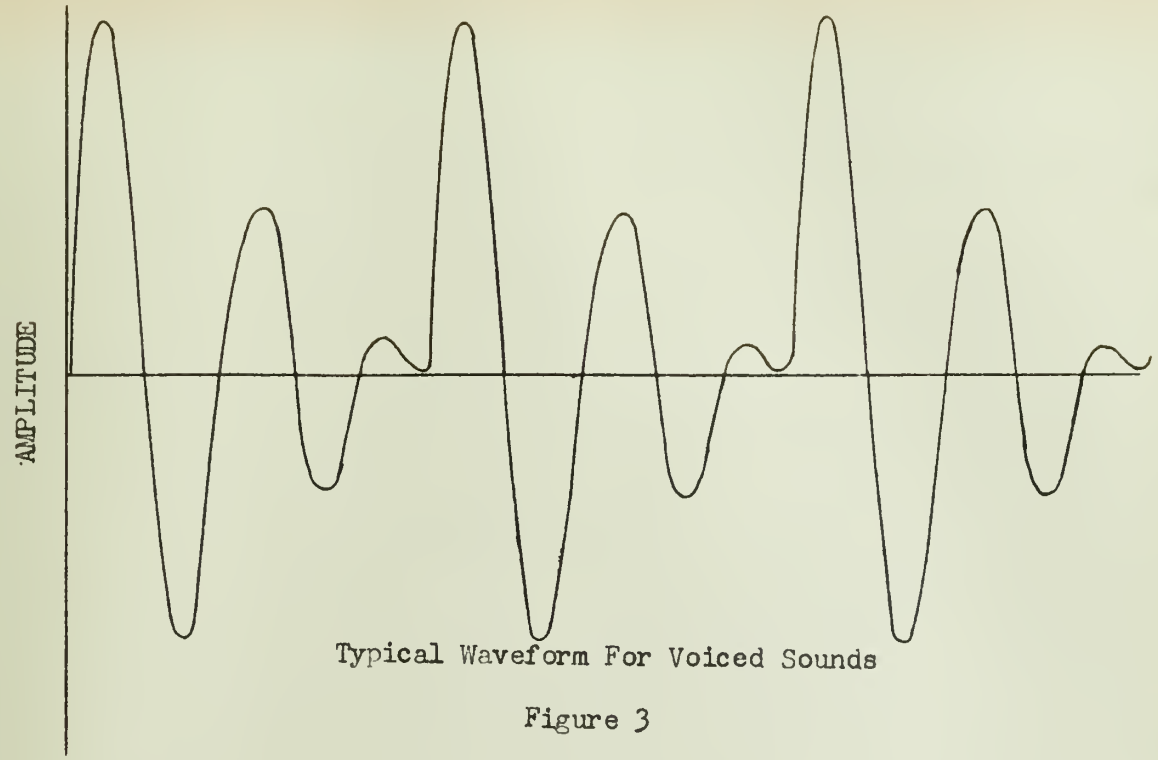
Pressure Waveshape at Larynx During Voiced Sounds

Figure 1



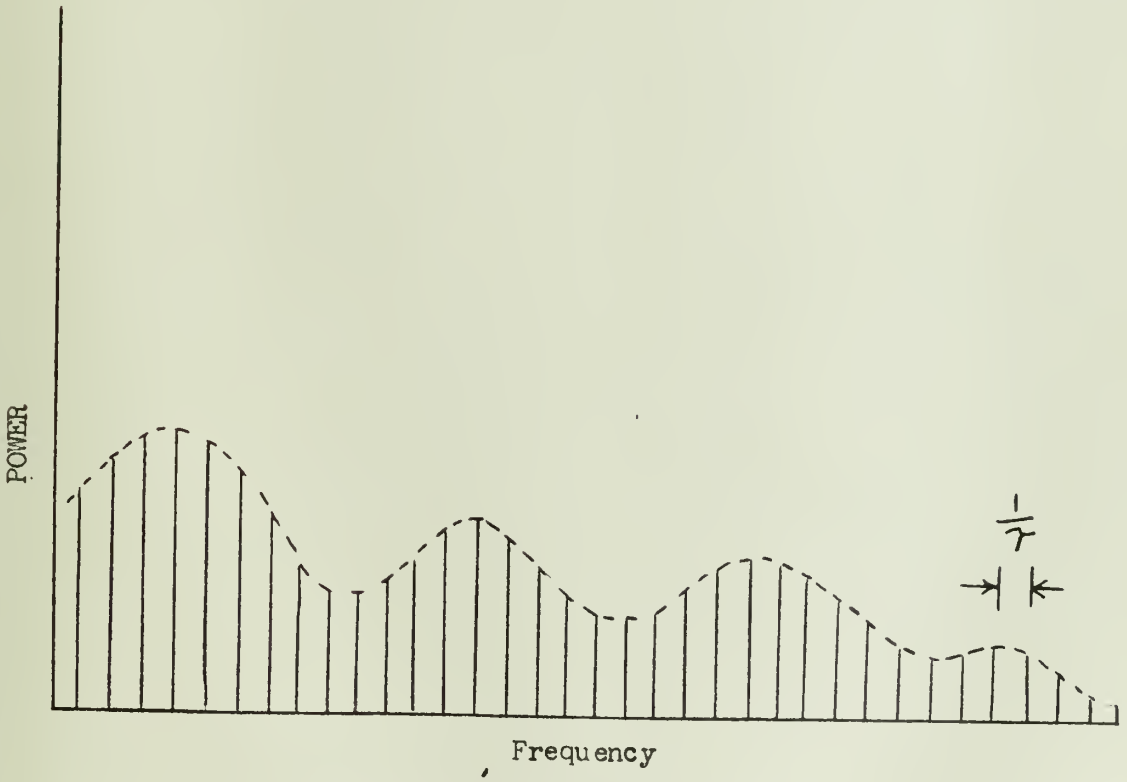
Approximate Spectrum of Larynx Source Energy During Voiced Sounds

Figure 2



Typical Waveform For Voiced Sounds

Figure 3



Typical Spectral Distribution for
Voiced Speech Signal Energy

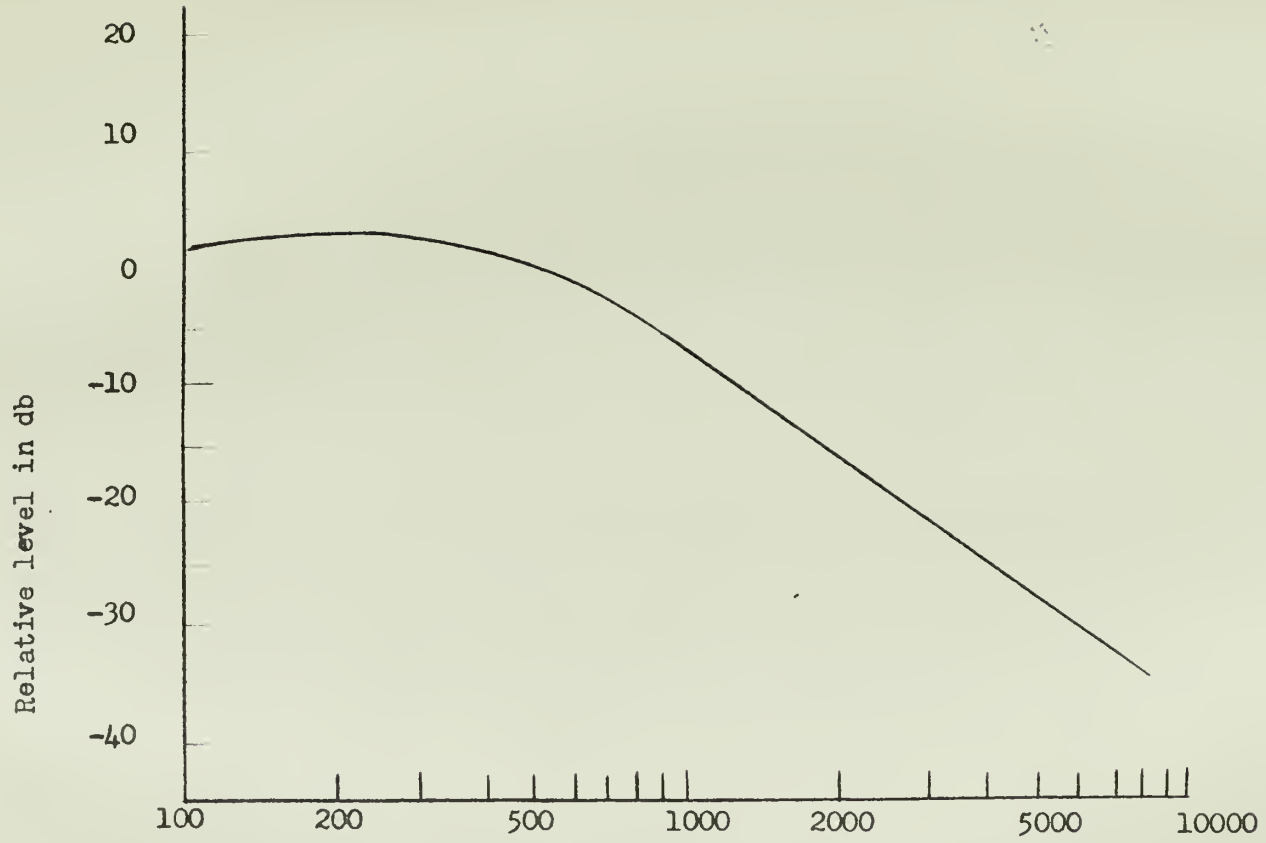
Figure 4

they shift around. Formant positions for a given sound for a given speaker also are not always in the same position. In general, the acoustic stimuli for a given sound and for speech varies from speaker to speaker and from a given speaker with time.

A typical long-time average of the voice spectrum is shown in Figure 5. A consideration of this curve shows that almost all of the power of speech is below 6000 cps. As a result, speech processing techniques have dealt with speech as if it were bandlimited to an upper value of 6000 cps. The telephone system has shown that a high degree of intelligence results when only one half this amount of bandwidth is considered. The effect of cutting off high and low frequencies on the articulation of different classes of speech has been investigated by Steinberg.[8] Figures 6, 7, and 8 show some of the results of his investigation. From these curves it appears that frequencies below 400 cps and frequencies above 6000 cps can be removed with little effect upon articulation.

Speech communication may be likened to a black box. The input to the box is the speech wave. At the output is the information perceived by the human sensor. Inside the black box is the auditory perception mechanism about which little is known. The goal of speech processing is to reduce the data in the speech wave by some scheme, present this reduced data to the input of the black box and have the human sensor perceive the same intelligence from the reduced data as he would if the input wave were the original speech wave. For this investigation the intelligence perceived has been defined to exclude such information as: 1. emotional status of the speaker; and 2. speaker recognition.

Experimentation on the inputs to the black box and observation of the intelligence perceived has lead to hints and clues about the nature of the auditory recognition mechanism. First of all, the auditory recog-



Typical Average Spectrum of a Voice Signal

Figure 5

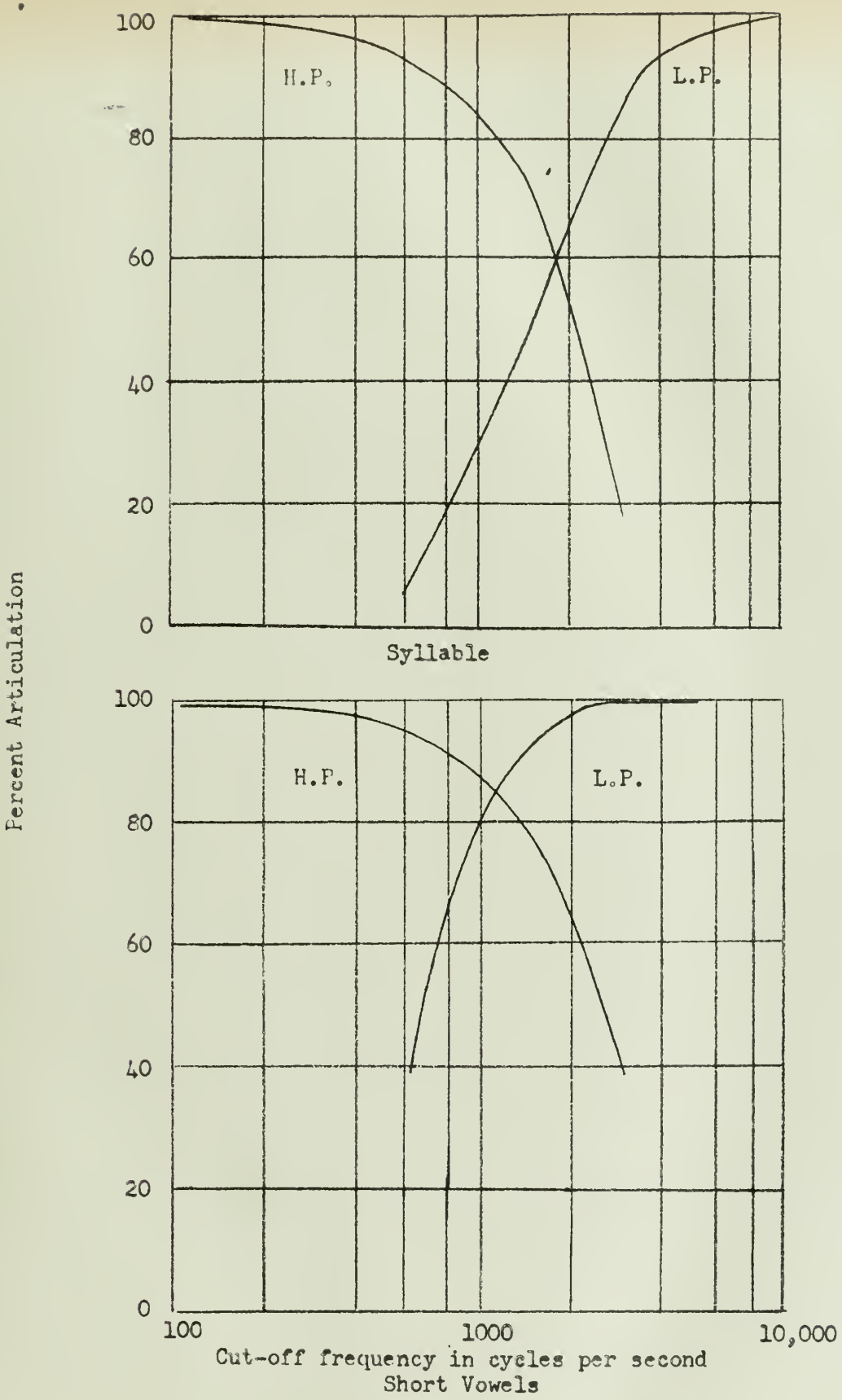


Figure 6. The effects of cutting off high and low frequencies on the articulation of different classes of speech sounds.

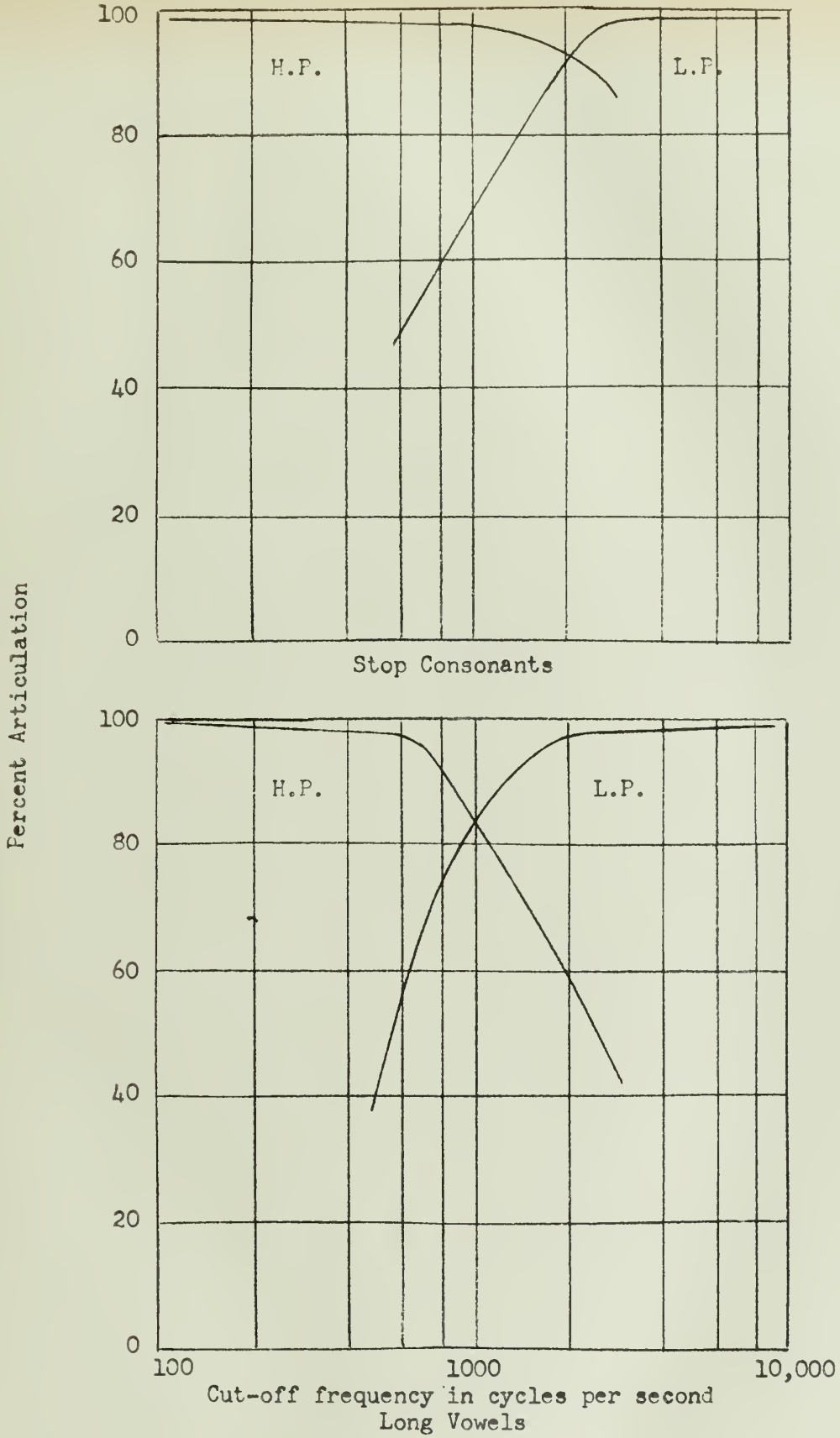


Figure 7. The effects of cutting off high and low frequencies on the articulation of different classes of speech sounds.

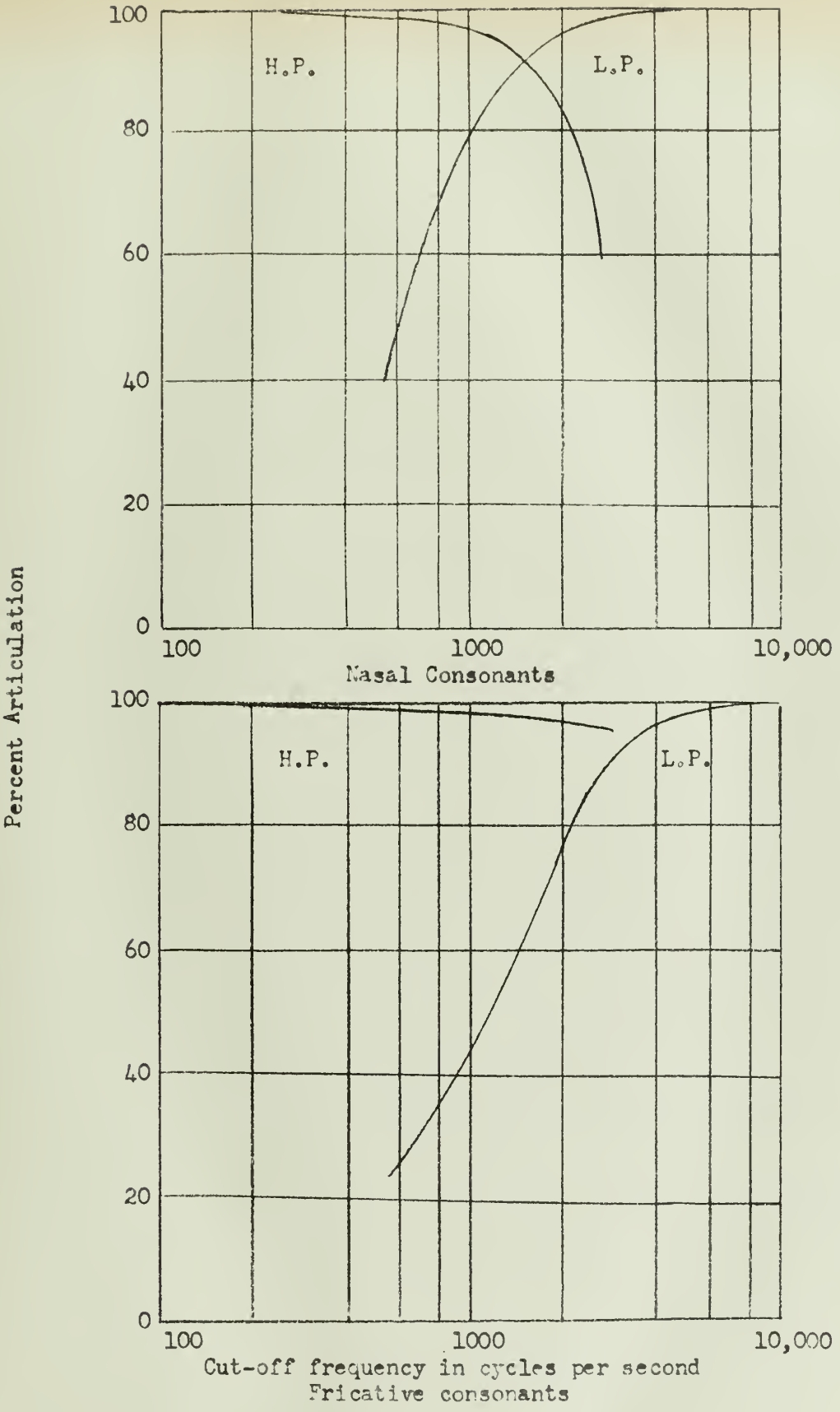


Figure 8. The effects of cutting off high and low frequencies on the articulation of different classes of speech sounds.

nition mechanism is not a constant parameter mechanism. If one doubles the frequency of a pure tone, the pitch perceived by a listener is not twice as high. Work by Stevens and Valkman[9] has led to the establishment of a pitch scale which relates the sensation caused by a frequency to the frequency producing it. Figure 9 shows the relation between pitch in mels and frequency. Similarly, the relationship between intensity and loudness of the acoustic stimuli in non-linear.

The human sensor frequently supplies information for which there appears to be no stimuli in the physical signal. If a listener is presented with a pure tone, he may report he also hears the harmonics.[10] In fact, if an auxiliary oscillator is introduced at a frequency three or more times the original tone, listeners also report they hear a beat frequency with one of the aural harmonics. The pitch heard from the original tone may also be varied by changing the stimulus time. If a listener hears a tone for 20 milliseconds, he will report that the pitch is lower than if he heard the same tone for five seconds. The shortest note which sets up any sensation of pitch has a duration of approximately 10 to 20 milliseconds.[4]

The human sensor is also capable of supplying the fundamental if only the harmonics are given. The frequencies 2000, 2200, and 2400 cps will separately cause perception of pure tones with a pitch of 2000, 2200, and 2400 cps. Together they will lead to the perception of a sharp sound with a pitch of 200 cps.[7]

Ohms law of hearing is frequently quoted as though the ear were absolutely insensitive to phase— this is not the case. The ear is relatively insensitive to phase and the phase angle may be varied only fairly wide limits, but an extremely wide variation will cause a change in the sensation perceived by the listener.[10] The ear is sensitive to the number and

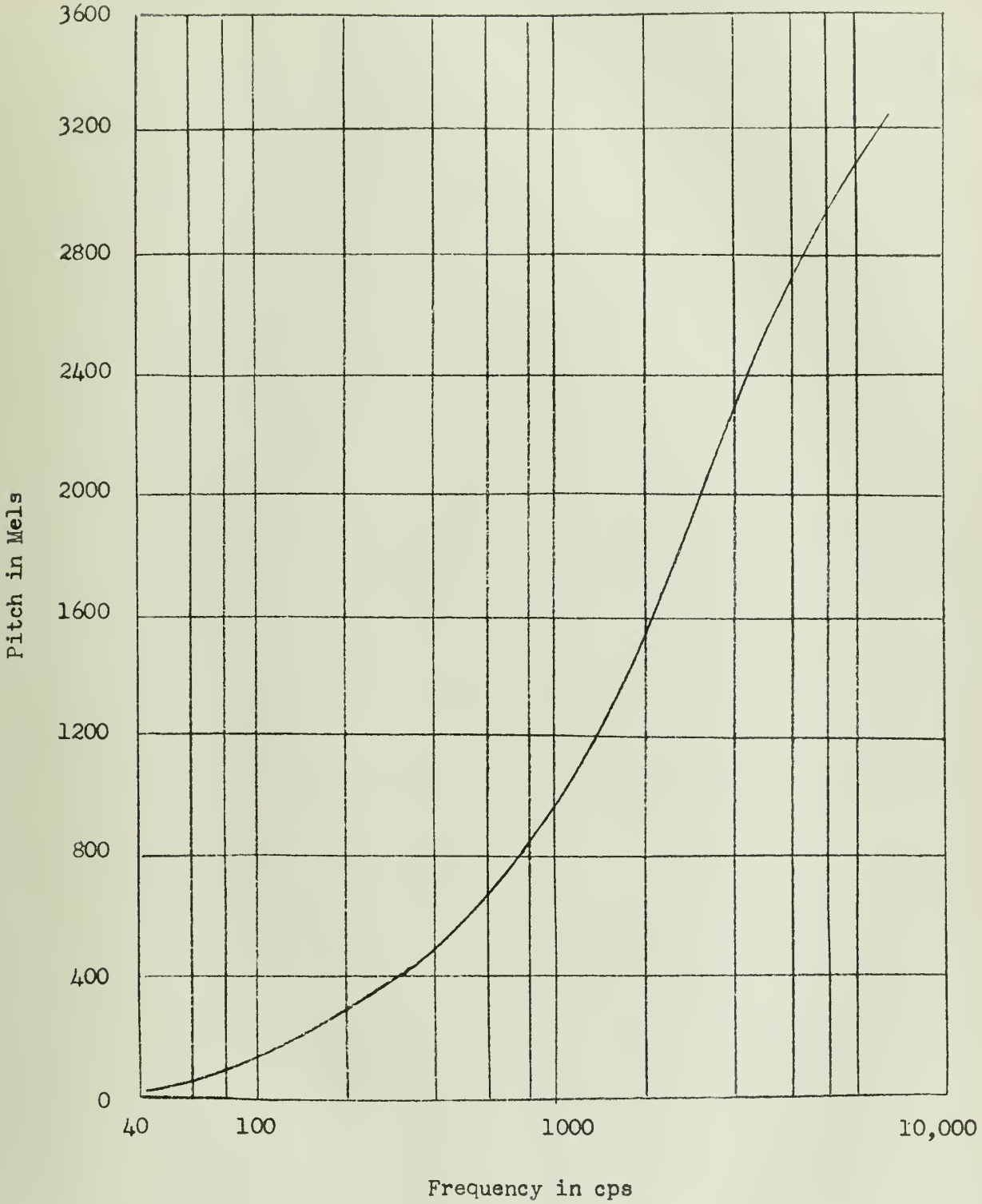


Figure 9. Relation between pitch and the frequency of pure tones 40 db above threshold.

amplitude of harmonics though; the quality of the sensation depends markedly upon the spectrum of the sound.

One of the most important characteristics of speech is its great redundancy. Speech processing techniques in general remove vast quantities of the information contained in speech and yet the artificially reconstructed speech is intelligible. The high degree of redundancy contained in speech has been demonstrated by a number of experiments. Licklider[11] has shown that up to 75% of the speech waveform, in the time domain, may be removed with practically no deterioration in intelligibility. Consider the high degree of redundancy involved if one can throw away 75% of the speech waveform and still have intelligibility. The success of the well known clipped speech systems in which the only information extracted from speech by the speech processor is the zero crossings along the time axis is indeed amazing and further points to the high degree of redundancy. There are a great number of speech processing schemes and the operation of each of them depends upon the great redundancy of speech. The success of these schemes in itself is a testimonial to this redundancy characteristic.

Another important factor that must be mentioned in connection with speech processing is that of a priori information. The a priori knowledge or psychological set of the human sensor with reference to auditory recognition is another factor which has enabled success in the speech processing field. The concept of psychological set is still very hypothetical and little understood today. Generally speaking the human sensor appears to possess a psychological set against which the incoming acoustic stimuli is compared to achieve intelligence and recognition. This concept is analogous to that in information theory in which we regard the receipt of signals as providing evidence of the messages selected at the transmitter, such

evidence converting the receiver hypthesis concerning the possible messages from an a priori set to an a posteriori set from which the receiver can make a best guess with a chance of error. The ability of the human sensor to fill in distorted or unrecognizable words in connected text has long been common knowledge. A possible explanation for this phenomenon is that the mind weights certain members of its psychological set on the basis of the subject being discussed and selects that member with the highest probability of occurrence when a word is missed. The tremendous ability of the mind to derive intelligence from only the barest hint of information is both a help and a hindrance to the communication engineer. But the help is major while the hindrance only minor.

The communication engineer in evaluating a speech processing system must determine to what factor any success of the system is attributable: the speech processing scheme itself or the tremendous ability of the human sensor. If a listener reports a high intelligibility score when connected text is used to evaluate a speech processing technique, doubt still remains about the actual performance of the processing scheme itself. A true evaluation must be based on an evaluation which uses isolated words; an evaluation in which the listener has no change to "pre-weight" certain members of his psychological set. A curve showing the relationship between word and sentence intelligibility is shown in Figure 10.

Before discussing the aid a priori information gives to speech processing, a few more general remarks about a priori information itself will be made. The psychological set of the human sensor is a product of his past environment. It seems fairly clear that the mind must store information about what words are expected to be connected with some concept, some idea, some topic of discussion. Similarly information must be stored about

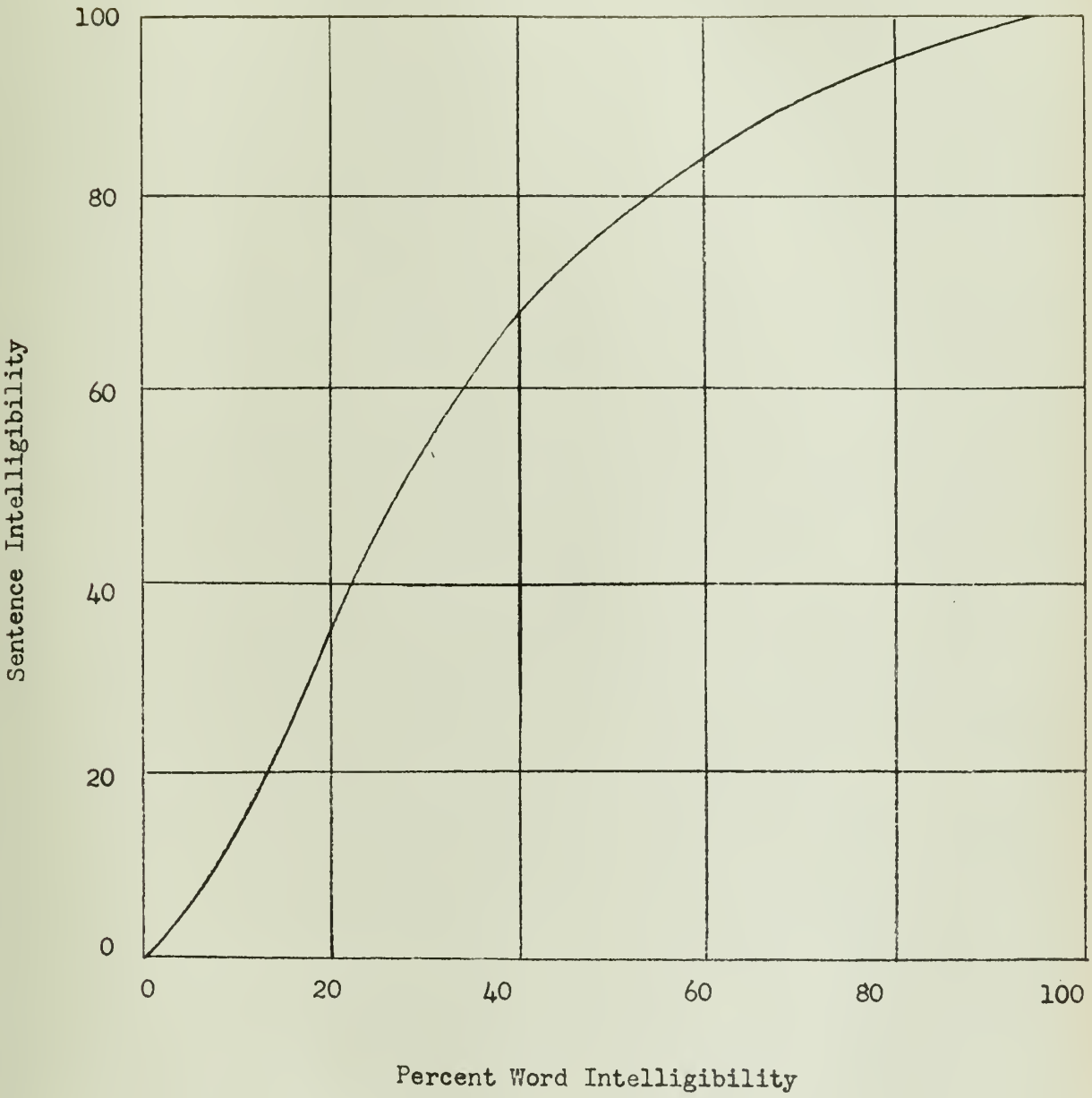


Figure 10. Word vs sentence intelligibility.

sentence structure, word groupings, and the set of expected acoustic stimuli from a given speaker. During conversation the listener weights certain members and subsets of his psychological set as determined by the current environment, recognition of the speaker, and subject matter. Thus, even before the listener hears the speech wave certain subsets have been essentially removed from consideration and the probability of correct recognition is enhanced. In a discussion about abstract art one certainly does not expect the interjection of a sentence about the social structure of an ant colony.

Most people at one time or another have talked with some person whose foreign accent was so thick that initially it was difficult to understand his words. But after listening to the speaker for some time one notices that it becomes easier and easier to understand him. The mind lacked a subset of expected acoustic patterns in this case and had to create a set before a high level of understanding was achievable. When the listener again meets this speaker and recognizes him by the sight mechanism it appears logical that the listener weighs the particular subset for the speaker and thus achieves more instant aural recognition.

D. B. Fry[12] has presented a demonstration of the manner in which a priori knowledge bears upon recognition. A phonograph record of the conversation of two speakers was distorted so that not a word of the conversation was recognized by a group of listeners. After the record was played once the listeners were told about the subject of discussion between the two speakers. When the record was played a second time most listeners were able to follow the entire conversation.

A priori knowledge may be reasonably expected to be a great aid in speech processing for listeners hearing the distorted artifical speech

of a processing scheme can build up a subset of expected sounds and thus bring about an enhancement of the success of the system. This phenomenon was observed during the investigation of the particular speech processing scheme described in this paper. After working with the system for a period of time it seemed obvious to the investigator that a certain group of sounds were indeed a certain word. But, if other listeners heard the same group of sounds for the first time, there was an element of doubt in their recognition. A discussion of speech processing and a priori information is presented by Cherry.[4]

4. Contemporary Speech Processing Systems.

Speech bandwidth reduction systems may in general be grouped into four principle categories:

1. Time or frequency compression methods.
2. Continuous analysis-synthesis methods.
3. Discrete sound analysis-synthesis methods.
4. Sound group analysis-synthesis methods.

Time or frequency compression systems utilize sampling or frequency division techniques. The Doppler Frequency Compressor system falls within this category.[13] One of the important forms of redundancy present in speech is repetition of the waveshape characteristic of a given sound during its generation. One could therefore obtain a 2:1 bandwidth reduction by: 1. sectioning the incoming speech wave into equal time sections; 2. transmitting only information on alternate sections; 3. reconstructing speech at the terminal end of the system by double playbacks on the information received on alternate sections. The Doppler compression scheme sections the incoming speech wave and then discards alternate sections. The remaining sections are expanded to twice theirnormal time interval thus filling out the blank time intervals generated above. The time expansion results in a compression of the frequency range of the sections to one half of its unexpanded value. The reduced data which has information spread continuously along the time axis is transmitted to the synthesizer which time compresses the incoming expanded sections to their original interval. This action expands the frequency range to its original limits and produces an alternating sequence of blank and signal filled intervals. Each signal filled interval is played twice by the synthesizer thus obtaining a continuous output. Experimental results indicate that compression

ratios of 1:4 to 1:6 may be achievable by this method. This system operates especially well with long vowel sounds in which the characteristic waveform is repeated many times. This scheme must be classed as one in which mild processing is accomplished, for at the synthesizer the alternate time compressed sections are exact replicas of the corresponding time intervals in the incoming speech wave except for spurious noises caused by the sampling mechanism.

David and McDonald[14] have developed another scheme utilizing time and frequency compression techniques. The techniques involve a pitch synchronous processing of speech. The feasibility demonstration of this technique involved two major processing steps, one of which should not be required in an operational system. In step one a channel vocoder was used to provide a convenient source of monotone speech for the input to the pitch synchronous analyzer. The pitch frequency for the monotone speech was set at and remained at 200 cps during the demonstration. The procedure of setting the pitch frequency was one of convenience and does not detract from the demonstrated feasibility of the system. As has been stated before, during voiced sounds there is a characteristic repetition of a basic waveform. The function of the pitch synchronous analyzer is to remove N-1 of these repetitions from the incoming speech and process the Nth period for transmission. The channel capacity required to accommodate only the information contained in the Nth period is thus 1/N of that required for the complete speech signal. The synthesizer reprocesses the information received on the Nth period to put it into the proper time and frequency frame, then plays the information once and repeats it N-1 times. Unfortunately, speech frequently contains sections which show little or no periodic structure. In the demonstration using monotone

speech as an input pitch synchronous processor the unperiodic sequences were segmented at the same rate as the voiced portions. In spite of this arbitrary sectioning of the unperiodic sounds, the resulting articulation was better than expected. In an operational system the scheme would not use a vocoder to provide monotone speech but would use the actual pitch frequency as a basis for segmentation. The treatment of the unvoiced sequences in an operational scheme still remains an unanswered question. Two proposals for their treatment have been made: 1.leave the unperiodic sections intact and code the information using an elastic time base to fit the transmission channel required for periodic information; and 2. segment the unperiodic sounds at some arbitrary rate. It is possible that there may be appreciable variation in the waveform between sampling intervals. In order to overcome this problem it has been proposed that the system, instead of repeating the one transmitted sample N-1 times, perform a linear interpolation at the synthesizer between adjacent transmitted samples; each such synthesized period is a step in the interpolation sequence. Experimentally it has been shown that for N as great as 6, using monotone speech as the input, the processing did not destroy the fundamental phoemic information.

The continuous analysis and synthesis schemes are those in which a number of analogue control signals are extracted from speech and transmitted to a synthesizer where they are used to control the operation of networks which are functional approximations of the human voice production mechanism. These control signals are associated with some parameter of speech and carry information about the activity of this parameter. For instance, a control signal may be associated with the amount of energy in a given frequency range of speech. Thus, for a high control signal

level there is associated a high energy level in the particular frequency band. There are a number of parameters of speech the rates of change of which are limited to syllabic rates of change.[2] The associated control signals carry information about the magnitude and thus the variation of these parameters. Since the chosen parameters vary at syllabic rates, about 15 to 25 cps, the control signals require a bandwidth of only 15 to 25 cps for transmission.

The goal of investigation in the continuous analysis-synthesis area has been and still is to judicially select to discover slowly varying parameters of speech, the utilization of which will lead to the reconstruction of satisfactory artificial speech with a minimum number of control signals.

There are a great number of continuous analysis-synthesis speech processing schemes. An adequate review of all of them is beyond the scope and purpose of this paper. A few of the more well known schemes will be discussed in order to point out current trends in this area and to serve as a background for the continuous analysis-synthesis scheme presented in this paper.

The Vocoder is perhaps the prime example of this type of scheme.[15] In this scheme speech is broken up into a number of contiguous frequency bands by an analyzer filter bank. The number of channels designates the type of vocoder: 12 channel vocoder, 18 channel vocoder. A group of analogue control signals which are associated with the amount of energy in each of the bands is derived by amplitude detecting the outputs of the analyzer filter bank. The control signals are transmitted to the synthesizer where they are used to amplitude modulate a local excitation function falling in a band corresponding to that from which they were derived. The

local excitation function at the synthesizer is composed of two types of excitation. One type of excitation is provided for voiced sounds, another type for unvoiced sounds. A buzz generator whose output is a harmonic spectrum with the fundamental and harmonics to a high degree provides excitation for voiced sounds. The fundamental of the buzz generator is controlled by what is called a pitch control signal. In the vocoder the pitch control signal is the output of a filter which passes frequencies from 100 to 300 cps in the speech spectrum. For unvoiced sounds a hiss generator provides broad and band noise excitation. The switching between energy sources from hiss to buzz is accomplished by the pitch control signal. When the speech is unvoiced there is no current in the pitch control channel and a switch in the synthesizer automatically switches in the hiss generator. A synthesizer filter bank identical to the analyzer filter bank receives the local excitation and breaks it up into channels identical frequencywise to the analyzer channels. Each channel in the synthesizer is then amplitude modulated by the control signal derived from the corresponding analyzer channel. The modulated signals from each band are mixed to produce the artificial speech. In essence, the system monitors only two types of parameters: 1. the energy in the given frequency bands; and 2. the lowest frequency present in the spectrum during voiced sounds. The associated energy control signal for each band sets the energy level for a corresponding band of excitation produced at the synthesizer.

In general, for satisfactory synthesized speech the number of channels has been between 10 and 18. The control signals vary at a rate of approximately 20 cps so that the bandwidth required for this system has been about 300 to 450 cps.

Stemming from the channel vocoder described above have been the for-

mant tracking vocoders two of which will be described.

In the resonance vocoder[16] speech is broken up into four channels; 40 to 400; 300 to 1100; 900 to 3000, and 3000 to 8000 cps. In each of the three upper channels, two parameters are monitored: 1. the total energy in the channel; and 2. the average number of zero crossings of the filtered wave taken over a finite interval. The pitch control signal is determined from the lowest channel in the same manner as the channel vocoder. The energy in the lowest channel is also monitored. The three upper channels are chosen such that they bracket the frequency regions in which the first three formants occur. It has been determined experimentally that the average channel frequency based upon the average number of zero crossings for each channel is a fairly good approximation of the formant frequencies F_1, F_2, and F_3.[17] Two types of excitation are provided in the synthesizer; buzz and hiss. The pitch control signal determines the fundamental of the buzz generator. The local excitation function is sent to three voltage variable resonant filters and to a 400 cps low pass filter. The frequency control signals associated with the formant channels adjust the center frequencies of the variable filters such that they correspond to the average frequency of each of the formant channels. The outputs of the variable filters are amplitude modulated by the associated energy control signals. The control signal associated with the pitch channel modulates the output of the 400 cps low pass filter. The type of excitation, buzz or hiss, is determined by a comparison between the energy control signals of the 40 to 4000 and 3000 to 8000 channels in the synthesizer. If the upper channel contains the most energy the hiss generator is switched in as the local excitation function. The operation

is completed with a mixing of all the modulated output in the synthesizer. It has been determined that for a total bandwidth of approximately 300 cps fair intelligibility results.

The second formant tracking vocoder to be described is a scheme developed by Howard in which seven parameters of speech are monitored.[17] The parameters extracted are the first and second formant frequencies, F_1 and F_2, their respective amplitudes, A_{F_1} and A_{F_2}, the voice pitch P the amplitude of the unvoiced turbulent sounds M_0, and the centroid of the turbulent sound spectrum M_1. The control signals associated with F_1 and F_2 are determined by averaging the zero crossing of the output of two voltage variable narrow bandpass filters. The center frequency of each filter is determined an auxiliary control signal which has been developed from the average number of zero crossings at the output of a fixed filter which br brackets the area in the speech spectrum in which the given formant occurs. A_{F1} and A_{F2} control signals are determined by envelope demodulating the outputs of the variable filters associated with formant frequency control signals. The control signal for M_1 is determined by averaging the zero crossing for the entire speech wave. M_0's control signal is derived from an envelope demodulation of the entire speech wave. The turbulent sound control signals are not transmitted to the synthesizer during voiced sounds. Turbulent sounds are synthesized by first amplitude modulating the output of a wide band noise generator with M_0 and then selecting out a portion of the noise spectrum with a voltage variable resonant filter whose center frequency is determined by M_1. Voiced sound synthesis is accomplished by: 1. feeding two voltage variable tuned filters in parallel with a series of short pulses the frequency of which is controlled by P; 2. adjusting

the center frequencies of the variable filters with the formant frequency control signals; and 3. amplitude modulating the outputs of the respective filters with A_{F_1} and A_{F_2} control signals. The modulated outputs of the turbulent and voiced sound synthesizers are mixed in the final step of the processing scheme.

Quantitative results for this scheme have not been published as yet. The estimated bandwidth for the scheme is approximately 140 cps for fair intelligibility.

Discrete and sound group analysis-synthesis methods will be treated together because the basic philosophy of the methods is the same. The methods differ only in the length of the sound group operated upon. The philosophy of these methods is to machine recognize a discrete sound unit and transmit a coded group identifying the unit to the synthesizer for voice reproduction. Synthesis may be accomplished by a simple readout of stored sound units from some memory device or a readout of a set of stored control signals to activate a speech synthesizer such as a vocoder. Phoneme recognition schemes operate on the sound unit with the smallest length. There are 40 phonemes utilized in the English language and a system which is capable of recognizing them would require only a 60 bit/second information rate to convey voiced information. To date there has been no successful demonstration of a device based upon phonemic coding.[6]

Investigations are also being conducted on methods which try to recognize groups of sounds that are composed of more than one phoneme but are shorter than a word. The use of pattern correlation matrices operating on sound spectrum shapes is the usual technique involved in the sound group schemes.

Recognition schemes which try to recognize entire words are at present

limited to very small libraries. These devices recognize only a few words and then only if the speaker for which the machine has been tuned speaks them.

5. A Speech Analysis and Synthesis Scheme for Bandwidth Compression.

The speech analysis and synthesis scheme investigated in this paper is a data-reduction scheme. The speech signal is destructively operated upon such that a high percentage of the redundant data in the speech is removed. The processed speech is then presented for transmission over a narrow band communication channel. The reduced data of the processed speech is used to control the speech synthesizer utilizing local excitation functions to reconstruct artificial speech at the terminal end of the system. The goal of this data reduction scheme is to achieve a bandwidth compression of the channel necessary to transmit speech information.

The scheme in question and the associated device break naturally into two areas: analysis of the complete speech waveform to achieve data reduction and synthesis of artificial speech.

The analyzer operates on a speech waveform to extract continuously seven low frequency coded signals as a function of time. These coded signals, which shall be called control signals, are a measure of seven parameters of the complete speech wave. It is the variation of these seven parameters that is important. Variations in the parameters are caused by changes in the articulation mechanism of a speaker and since these articulation changes are restricted to low frequency syllabic rates the channel width required for a transmission of each of the seven parameters is approximately 20 cps.[2, 17]

The major increase in efficiency comes from sending not the complete speech waveform which is complex but only information to control local excitation functions at the synthesizer. The data transmitted consists of how the speech is varying and is not speech itself.

The synthesizer using the incoming control signals to modulate local

excitation functions, similar in general to the physical sources producing the speech, reconstructs a representation of the analyzed speech thus producing artificial speech.

The functional block diagram for the speech analyzer is shown in Figure 11. From this diagram it is seen that the seven control signals extracted from the complete speech wave may be divided into two basic types. Three control signals consist of amplitude information; four consist of frequency information.

The scheme extracts frequency and amplitude information from the same regions in the speech spectra with the note that amplitude information for the pitch channel is not extracted from the frequency region normally associated with pitch. For example, observe that both amplitude and frequency information are extracted from the region 3000-6000 cps.

Investigations carried out in an allied speech-processing area by W. C. Dersch at IBM, data yet unpublished, tend to indicate that the optimum area to extract amplitude information may not necessarily be the same as the area of extraction of frequency information. Nor, does a number of frequency extractors have to be the same as amplitude extractors. Further extensive investigation is required to optimize the number and placement of the frequency and amplitude extractors in the voice spectrum.

The incoming speech waveform upon entering the analyzer is separated by fixed filters into three frequency bands; 300-1500 cps, 1500-3000 cps, and 3000-6000 cps. The outputs of the various filters are sent to the frequency and amplitude extractors associated with that particular channel. The output of the 300-1500 cps filter is also sent to the pitch extractor circuit.

The function of the amplitude extractors is to derive an indication

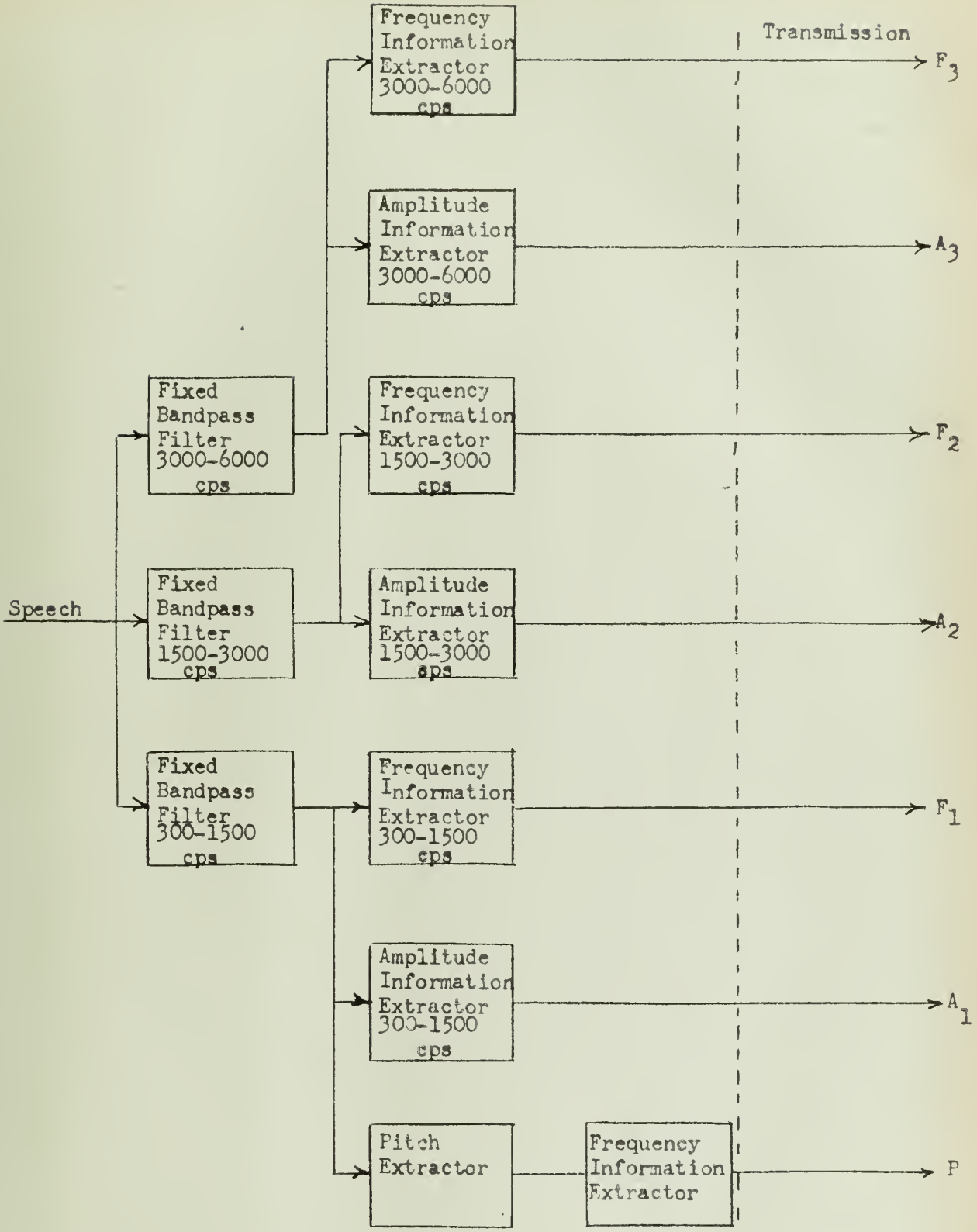


Figure 11. Functional block diagram of the speech analyzer showing the development of the seven control signals to be transmitted to the speech synthesizer.

of the energy present in the various major bands as a function of time. A graphical presentation of the output of the amplitude extractors is also shown in Figure 12. The amplitude extractor takes the signal emanating from its associated fixed filter, envelope demodulates it, smooths the resulting waveform, and filters the output to allow only variations of approximately 20 cps or below. The circuitry to accomplish these functions is shown in Section 6. Due to the smoothing action of the demodulator and filter the resulting control signal cannot be said to be an absolute instantaneous measurement of the energy in the band. It is a very close approximation.

A complex waveform, over a given time interval, may be completely specified with a Fourier series. Information theory has shown that a waveform may be completely specified with the correct number of discreet samples during a given time. An approximation, it must be admitted gross, to a waveform is obtained if one specifies only the axis crossings, zero crossings, of the waveform and assumes that the waveform is sinusoidal in nature between the zero crossings. This approach is, of course, the clipped speech approach as discussed by Licklider.[11] Consider the extreme destruction performed on the speech wave when only the zero crossings of the wave are transmitted. The surprisingly high intelligibility resulting when tilting and differentiation are performed prior to clipping is indeed factual evidence of the great redundancey of speech and the small amount of information that must be presented to the human sensor for auditory recognition. A communication system, Frena,[1] has been developed in which the zero crossings and envelope of the speech waveform are transmitted. This device uses the resulting data reduction to obtain an increased signal-to-noise ratio rather than bandwidth compression.

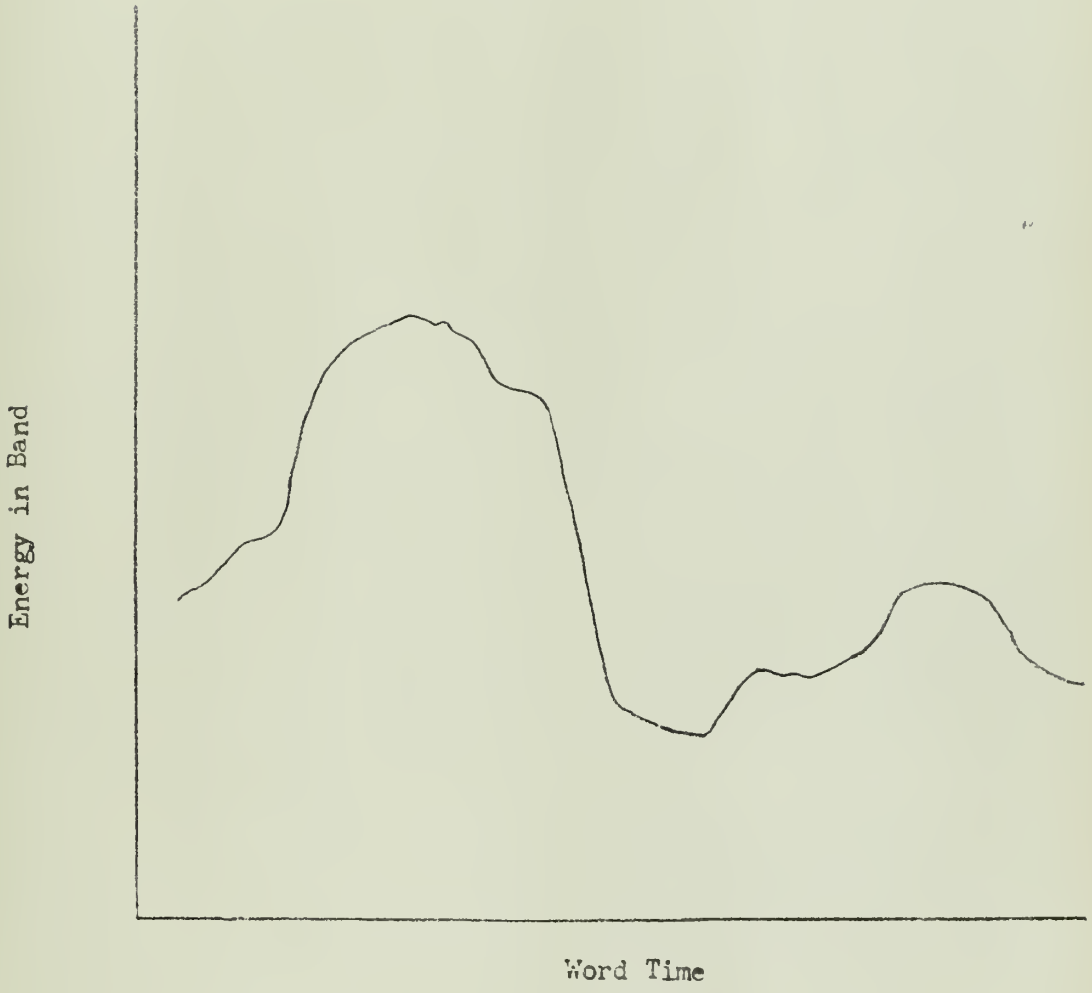


Figure 12. Typical output waveform of Amplitude Information Extractors.

This investigation has taken the approach that another type of waveform approximation is obtained if slope reversal information on a complex wave is utilized. The function of the frequency extractors is to derive from the complex wave at the output of the fixed filters a measure of the average frequency of the complex wave over a delta interval. This measure being defined as a short term average of the slope reversals of the complex wave.

The frequency extractors develop a pulse of given width each time the input wave reverses slope. An integrator operates upon the incoming pulse stream and produces a varying DC voltage which is a measure of the short time average of the slope reversals. Since the components that produce a change in slope reversal rates are limited to syllabic rates then the output of the frequency extractors will possess variations of the order of 20 cps. The control voltage produced by the frequency extractors, as has been stated, is a measure of the average frequency of the input waveform over a delta interval. The integration time of the frequency extractors is approximately 50 msec. Note that the slope reversal information is obtained from the output of the fixed filters and not from the complete speech waveform. Figure 13 shows a graphical presentation of the output of the frequency extractors.

Pitch frequency information is extracted from the frequency band 300-1500 cps. This is a radical change from the usual method of pitch frequency information extraction. The usual approach has been to use a band pass filter in the region from 100 to 200 cps to extract the fundamental of the Fourier series of the speech waveform and call this the pitch frequency.[5,6,15,17]

The frequency corresponding to the pitch of the male voice is in

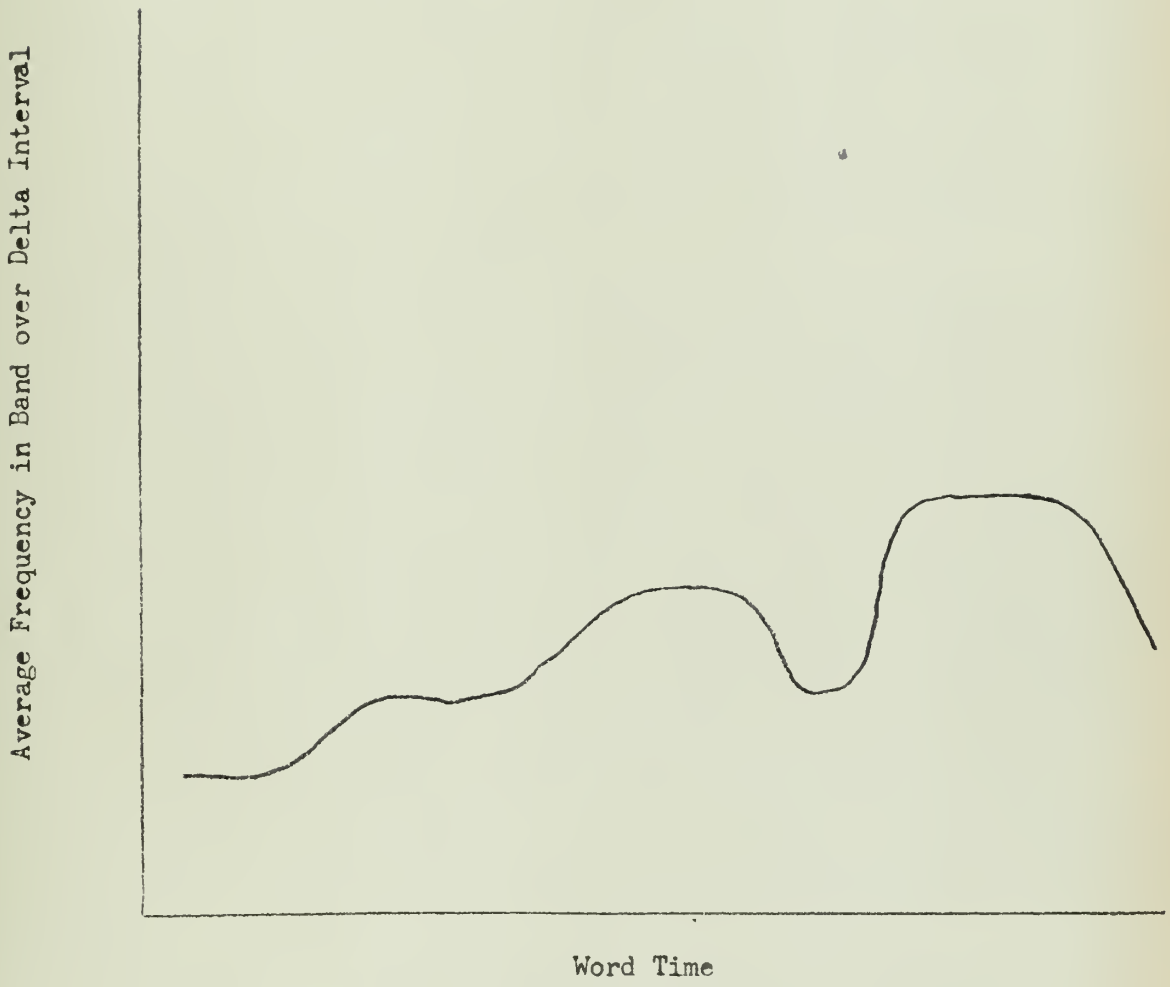


Figure 13. Typical output waveform of Frequency Information Extractors.

general below 200 cps. For female voices it may range as high as 500 cps. An interesting phenomena is that the human sensor perceives pitch regardless of whether a frequency corresponding to the pitch is present in the speech spectrum or not. Consider the telephone. All frequencies below 300 cps are not passed by the system. Yet, the listener hears pitch. Spectral analysis of speech waveforms has shown that very often the frequency corresponding to the pitch is not present in the speech spectrum or present to a very diminished degree.[7] Partially deaf persons who are deaf to all frequencies below 1000 cps, still in voice conversation distinguish pitch.

The view of pitch taken in this investigation is based upon the theory of the residue[7] which shall be discussed.

Consider first the inadequacy of the system which tries to extract pitch by filtering out the fundamental of the Fourier series which at a various times is not even present in the voice spectrum. No amount of filtering is going to extract a frequency that is not present. A cognizance of this problem has resulted in fundamental "finders" which are complicated and often not much more proficient than the approach of finding the fundamental by filtering.[18,19,20] These "finders" in general attempt to track two harmonics in the speech spectrum and from these harmonics obtain a beat frequency corresponding to the fundamental. Unfortunately, sometimes the particular harmonics being tracked absent themselves from the spectrum.

In general the frequency corresponding to the pitch as perceived by the human sensor is the fundamental of the Fourier series.[7] But, sometimes it is not.

Considering the illustrations of the telephone, spectral analysis,

and partially deaf persons, then by what means does the human senor perceive pitch when the acoustic stimuli does not contain a frequency corresponding to the pitch? The residue theory contends that a collective observation of the higher harmonics of the speech spectra results in the perception of a sharp sound, this sound component being called the residue. The collective vibration form of these harmonics is periodic in nature. The periodicity of the collective waveform, which is very apparent in the speech waveforms, corresponds frequencywise to the frequency of the residue. The periodicity of the collective waveform and the frequency of the residue corresponds almost all the time to the fundamental of the speech spectrum. In the remaining cases the waveform periodicity and residue frequency correspond to lower harmonic frequencies; i.e., second or third. In all cases, the frequency of the pitch perceived by the human sensor is the residue frequency.

Based upon the residue theory, the method utilized in this investigation to determine a measure of the pitch frequency is as follows. The pitch extractor monitors the output of the lowest frequency band fixed filter; that is, 300 to 1500 cps. During voiced speech the collective waveform of the harmonics in the band 300 to 1500 cps is periodic. The pitch extractor develops a sinusoidal waveform whose frequency corresponds to the periodicity of the speech waveform in this band. It has been found unnecessary to observe the complete speech spectrum; the periodicity of the unfiltered speech waveform being the same as the periodicity in the band from 300-1500 cps. The pitch extractor is composed of an envelope demodulator and a low pass filter network. The circuitry is shown in Section 6.

The output of the pitch extractor is sent to a frequency extractor

circuit which develops a control voltage which is a measure of the average frequency of the sinusoidal waveform at the output of the pitch extractor over a delta interval. Figure 14 is a series of photographs of the waveform at the output of the 300-1500 fixed filter, and the resulting sinusoidal wave at the output of the pitch extractor for three voiced sounds.

The functional block diagram for the speech synthesizer is shown in Figure 15. The function of the speech synthesizer is to utilize the seven incoming control signals to continuously synthesize speech.

The frequency information control signals operate to select the position of the passband in four voltage variable filters. The action of the voltage variable filter has been quantized. For example, when the frequency control signal for the sub-band 300-1500 cps varies continuously from a voltage that corresponds to 300 cps to a voltage that corresponds to 1500 cps, the center frequency of the passband of the associated voltage variable filter does not move continuously from 300 to 1500 cps but moves discreetly in a series of seven steps. Thus, the center frequency of the filter remains at 300 cps for control signal values corresponding to frequencies of 300 to 400 cps. At 400 cps the passband center shifts to 500 cps and remains there until the control signal reaches a value corresponding to 600 cps. This procedure is followed in all of the voltage variable filters. The filter shifts from one center frequency to the next at a frequency which is midway between the quantized filter center positions. Table 1 lists the quantized center frequency positions of the voltage variable filters. The passband of the four filters are: 20 cps for the sub-band 100 to 200 cps; 200 cps for the sub-bands 300 to 1500 cps and 1500 to 3000 cps; and 300 cps for the sub-band 3000-6000 cps. Different passbands were used in the various sub-bands for two reasons. First,

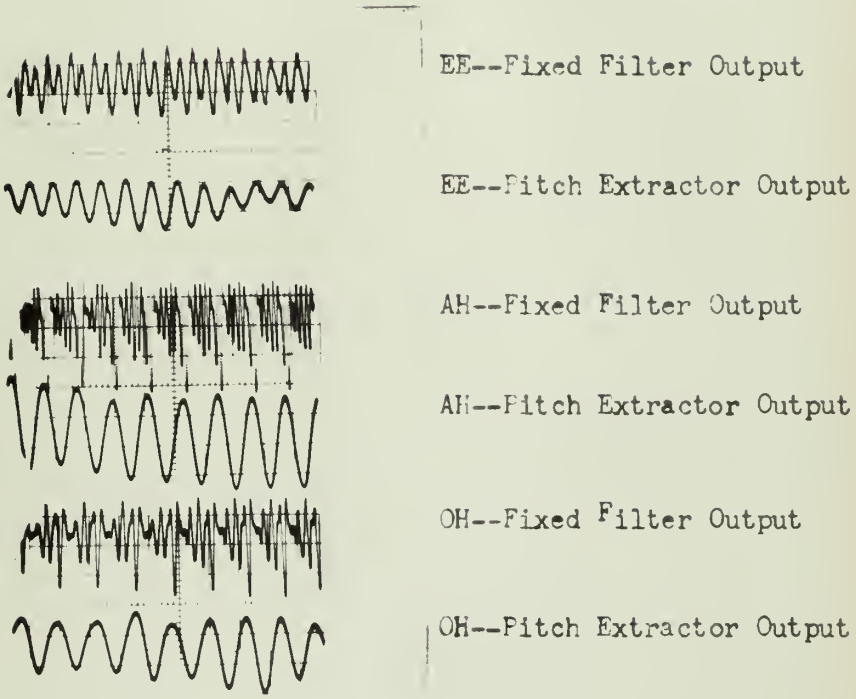


Figure 14. Output of fixed bandpass filter, 300 to 1500 cps, and corresponding output of Pitch Extractor for three voiced sound inputs: EE, AH, and OH.

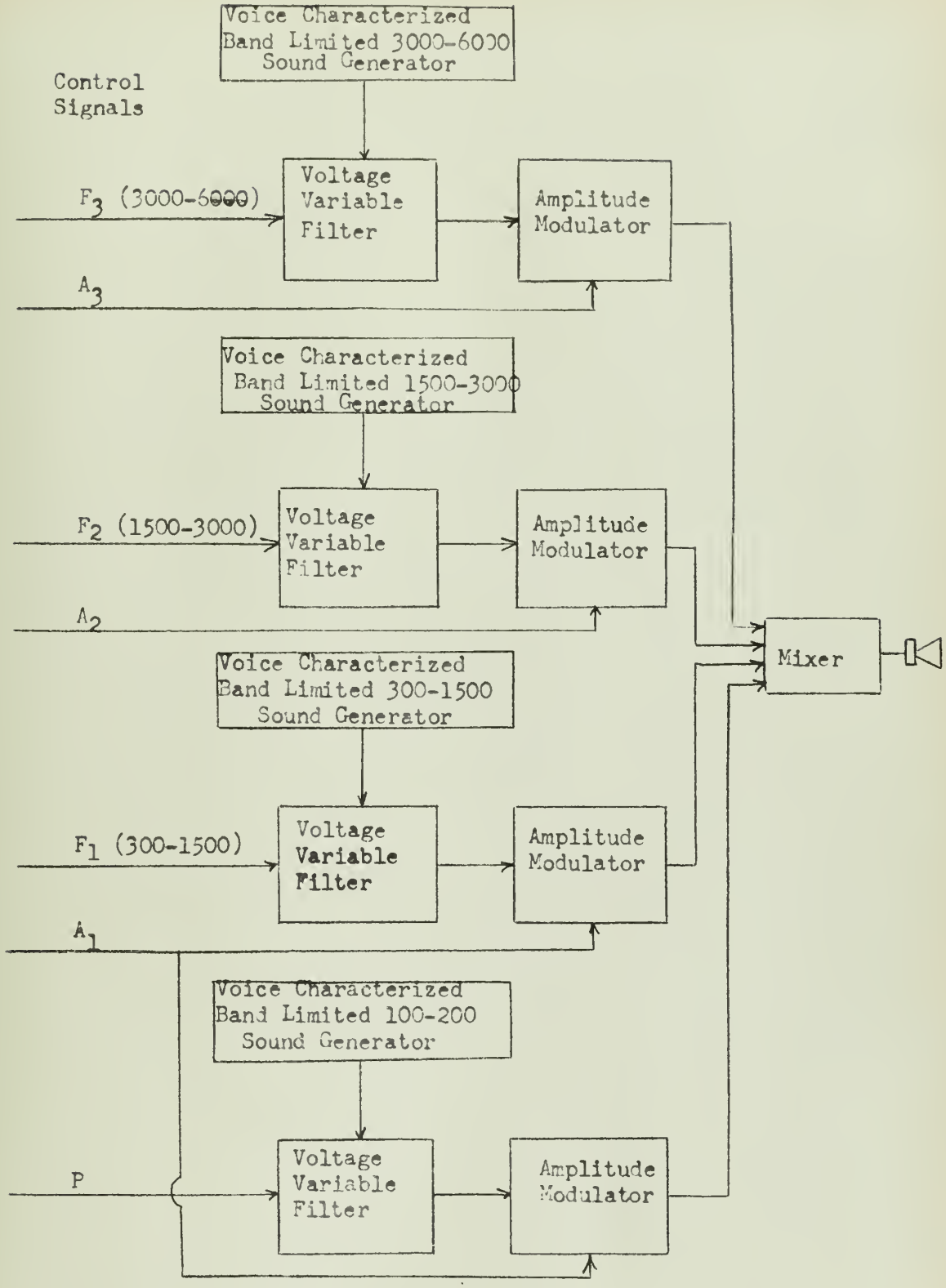


Figure 15. Functional block diagram of the speech synthesizer showing seven input control signals and speaker output of artificial speech.

the type of information connected with the lower sub-band is narrower than the type present in the upper sub-band. The lower sub-band is concerned with pitch information the upper sub-band involves mainly wide band fricative information. Second, the ability of the ear to differentiate between frequencies becomes poorer as frequency increases.

Each voltage variable filter filters the output of its own associated unique sound generator, the filter position being determined by the corresponding frequency control signal. A survey of the literature has shown that in other speech synthesis schemes functionally comparable sound generators are almost always in the form of buzz and hiss generators or oscillators. The operation of these devices is well understood. Here, instead of presenting to the filter the band limited white noise of hiss generators or the harmonically rich output of the buzz generators and oscillators, the approach taken is to present to the filter band limited voice characterized sound. The actual implementation of the sound generators may procede along a number of approaches. Tracks on a magnetic drum may be utilized. A single continuous groove on a phonograph record may be used. The method used in the investigation was to pass a single continuous loop of magnetic tape through a tape recording device. There were four tracks on the tape. Each of the four tracks is associated with a major frequency band in the synthesis scheme. That is, one track is associated with the band 3000 to 6000 cps, one with the band 1500 to 3000 cps, one with the band 300 to 1500 cps and one with the band 100 to 200 cps. The sound on each track is the result of a person or groups of persons speaking through a fixed bandpass filter whose limits correspond to the frequencies mentioned just above. Recording is done at an unsaturated level. After several cycles a track on the continuous tape loop is over recorded

many times and is thus saturated with band limited, voice characterized sound. This sound is not pure noise, but is sound which has the speech characteristics of the selected channel. The sound generator thus possesses characteristics of the human voice production device.

The outputs of the sound generators are one of the two inputs to the voltage variable filters. The other input to each of the variable filters is the frequency control signal associated with that channel.

The use of these sound generators is indeed empirical. It is a hypothesis of this scheme that the use of band limited, voice characterized sound will lead to increased intelligibility and naturalness in the synthesized speech. Research on speech sounds themselves has shown that the use of superposed samples results in a sound which displays the average spectral properties of speech more readily than the methods that have been employed.[21]

The outputs of each of the variable filters is amplitude modulated by its associated amplitude control signal.

It will be recalled that the frequency corresponding to the pitch was determined by observations on waveform periodicity in the lower frequency band channel. This frequency, in general, for male voices is between 100 and 200 cps so that while there is no analysis done on speech in the 100 to 200 cps region there must be a sound generator and variable filter in the synthesizer for this region in order to synthesize the pitch sound. The output of the pitch channel variable filter is amplitude modulated by the amplitude control signal of the 300 to 1500 cps channel. This amplitude control modulates the output of the variable filter associated with the 300 to 1500 cps channel. This illustrates the concept discussed earlier that the frequency and amplitude channels must not necessarily cover

the same range in the voice spectrum.

The outputs of the modulators are resitively mixed, amplified, and passed to the output speaker, the synthesis of artificial speech being complete.

The system block diagram is shown in Figure 16.

Figures 17-20 are photographs of oscilliscope presentations at various points throughout the system for the word "six". Figure 17 shows the input waveform to the system and the synthesized output waveform. Figure 18 shows the output of the three fixed analyzer filters. Figure 19 shows the four associated frequency control signals. The corresponding amplitude control signals are shown in Figure 20.

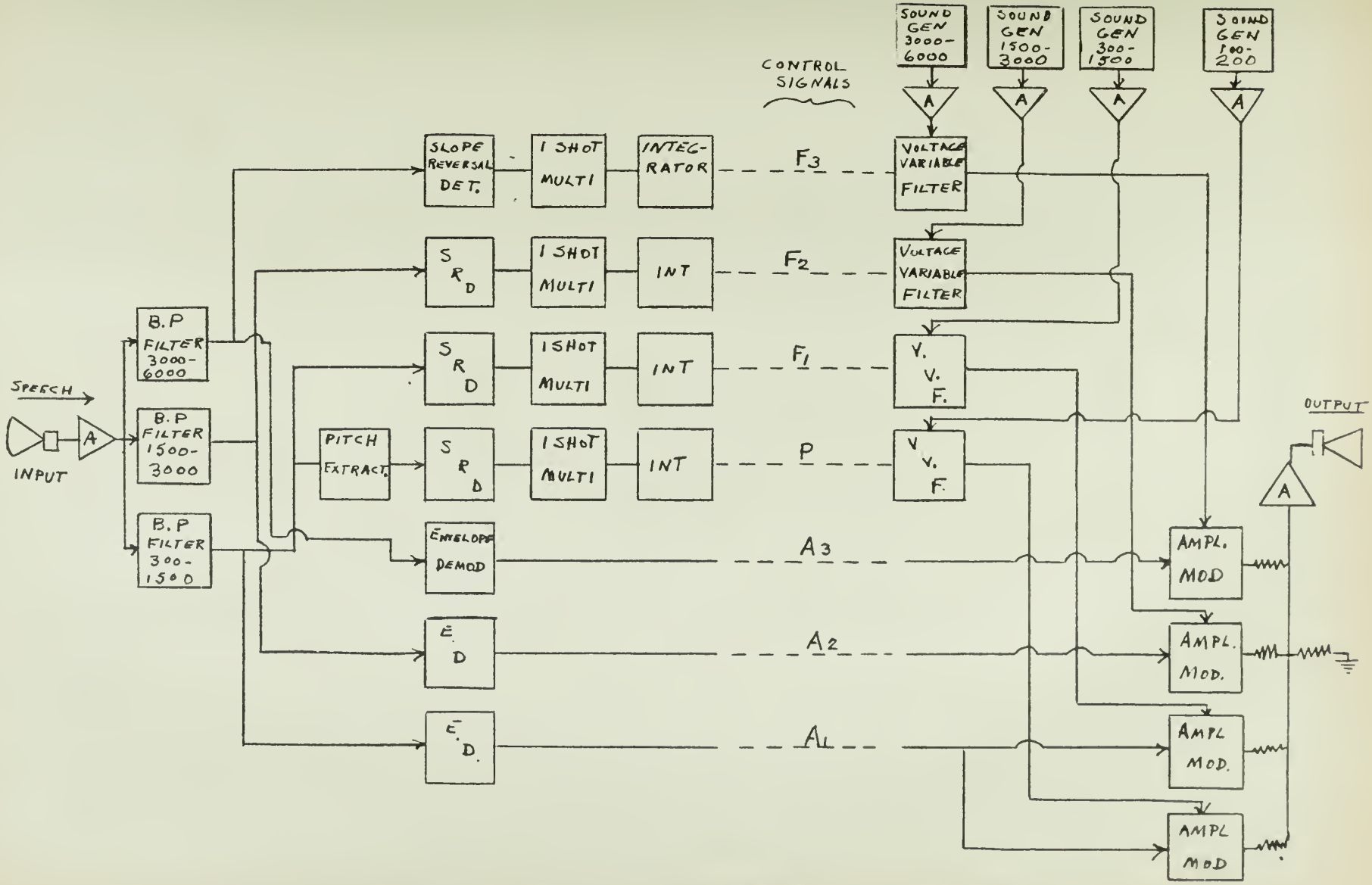


Figure 16. System Block Diagram for Speech Analysis-Synthesis Scheme

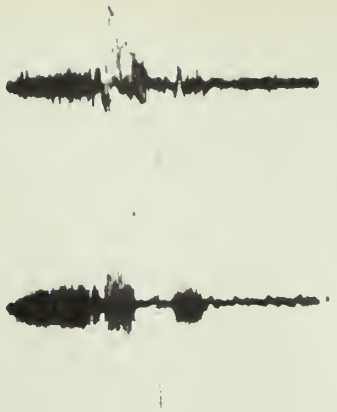
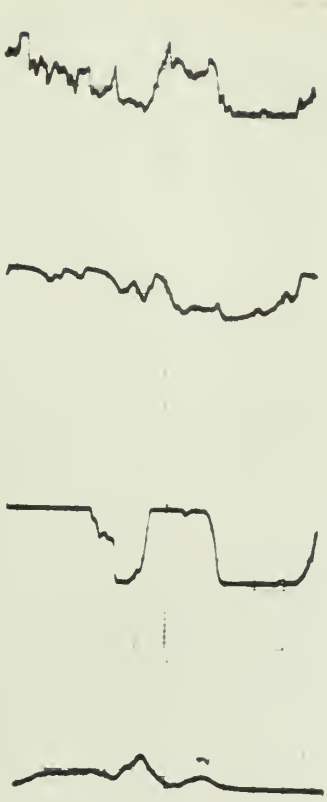


Figure 17. Top: Audio output waveform of system synthesizer for input word "six".

Bottom: Audio input waveform to system, for word "six".

Figure 18. Output waveforms of analyzer fixed filters for word "six". Top, 3000 to 6000 cps band. Middle, 1500 to 3000 cps band. Botton, 300 to 1500 cps band.



TIME SCALE 100 MSEC/CM

Figure 19. Frequency control signals for word "six"
From top to bottom:
1. 300 to 1500 cps band.
2. Pitch control signal.
3. 1500 to 3000 cps band.
4. 3000 to 6000 cps band.

TIME SCALE 100 MSEC/CM

Figure 20. Amplitude control signals for word "six".
Top to bottom:
1. 3000 to 6000 cps band.
2. 1500 to 3000 cps band.
3. 300 to 1500 cps band.
Note in the 3000 to 6000 cps band the buildup of energy during the "s" sounds and the drop-off during the voiced "i" sound. In the 300 to 1500 cps band observe the lack of energy in the band for all sounds except the voiced "i" sound.

6. Implementation of Speech Processing Scheme

The design level set during the investigation was based upon three philosophies. First, a degree of looseness is permitted and normal for the investigation and demonstration of a concept at the laboratory level. Second, the gap between the laboratory device and a functionally equivalent commercial product should be kept at a minimum and be easily traversed by simple product engineering. Third, when an element normally not associated with a given function is utilized, intensive design research and a more tightly engineered component is demanded in order to evaluate it both from a device and system standpoint.

The third philosophy characterized the voltage variable bandpass filter utilized in the speech processing system. The requirements set for this device were found to be higher than those currently being observed by investigators in closely allied speech processing research. The voltage variable filter is considered to be a key element in the system and as such had greater demands placed upon it. Much consideration was given during the design stage to the possibility that an inverse relationship might exist between system intelligibility and filter performance. Because of the critical nature of the bandpass filter a great deal of effort and time was spent in the choise of a circuit and its development. As a result, the treatment of the voltage variable filter is far more extensive than for other system components.

The design and construction of the various functional components of the speech analysis and synthesis system was, for the main part, straightforward. The finalized circuits for the more straightforward components shall be presented and discussed only briefly. A more intensive discussion will be presented for those components which posed a more serious

problem.

Referring to Fig. 16, we see that the speech waveform after passing through the microphone, is passed through a voltage amplifier. This voltage amplifier is of standard design. The output of the voltage amplifier is then sent to three SKL Model 302 filters. The bandwidth and center frequency of each of the pass bands may be varied by manual adjustment.

The circuitry for the amplitude information extractor is shown in Fig. 21. It consists of a standard envelope demodulator, a half-wave rectifier, followed by three low-pass filters. The low-pass filters perform two functions. They smooth the wave form and permit only variations of 20 cps or below. The output is from an emitter follower.

Fig. 22 shows the circuitry of the frequency information extractor. The frequency extractor is composed of three sections: a slope reversal detector, a monostable multivibrator, and an integrator. The slope reversal detector developes a trigger pulse for the multivibrator each time the input wave reverses slope from negative to positive. The multivibrator emits a train of constant width pulses which are short-term averaged by the integrator. The integration time is approximately 50 milliseconds. The circuitry shown in Fig. 22 is for the frequency band from 1500 to 3000 cps. The RC time constants of the multivibrator and integrator must be varied slightly to accommodate the other major sub-bands. A picture of the four frequency information extractors is shown in Fig. 23.

The pitch extractor shown in Fig. 24 consists of an envelope demodulator followed by two constant k low-pass filters. The function of the pitch extractor being to develop a sinusoidal wave whose frequency corresponds to the periodicity or pitch frequency of the speech wave for

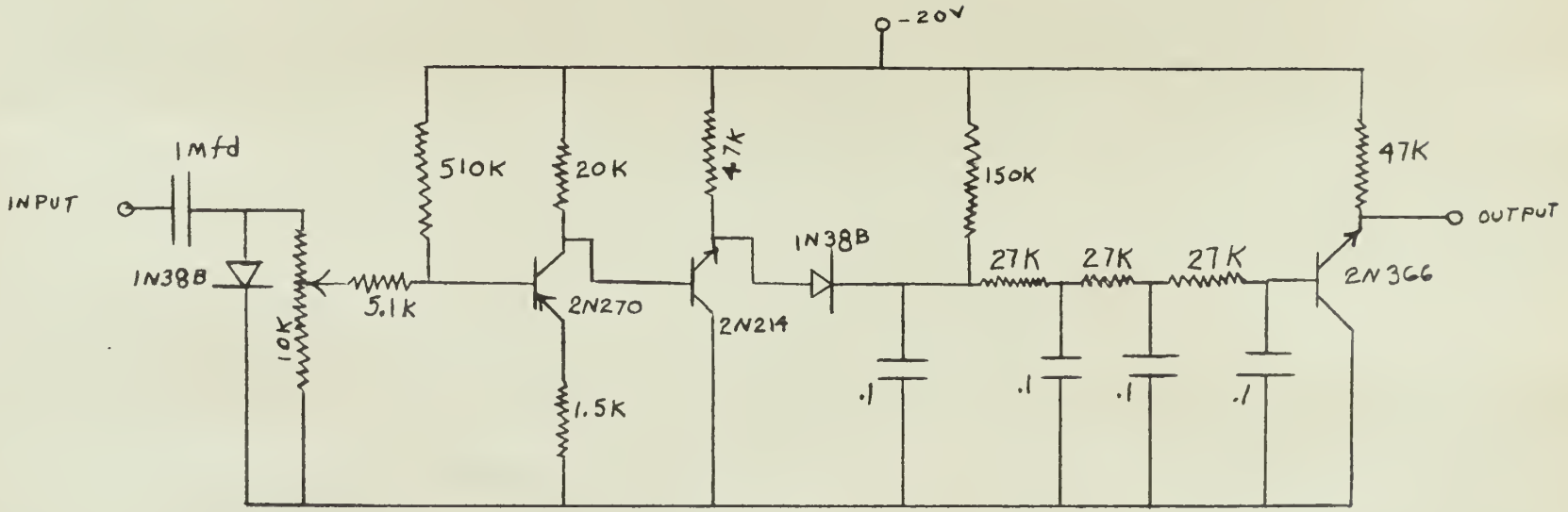


FIGURE 21. AMPLITUDE INFORMATION EXTRACTOR

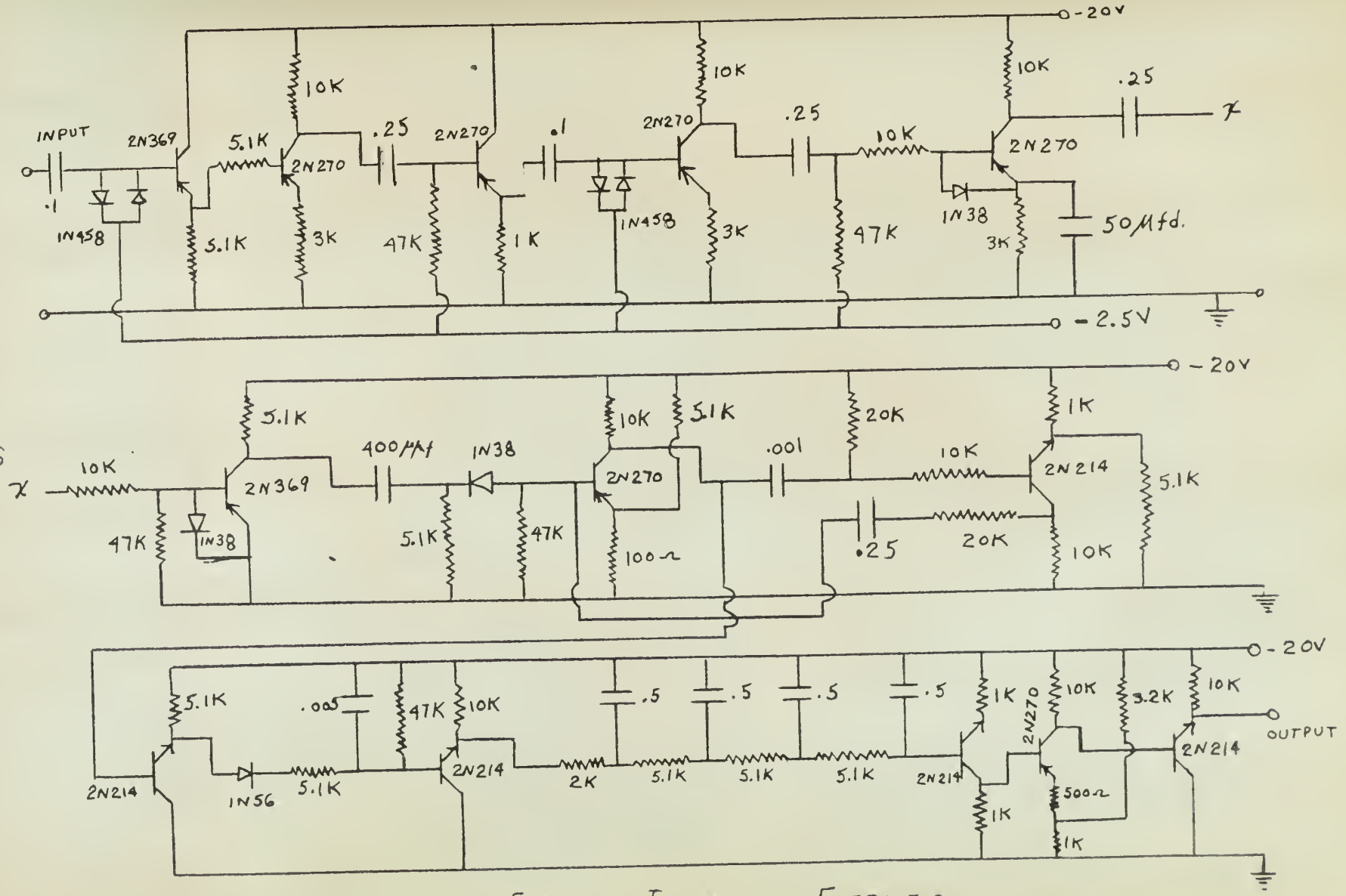


FIGURE 22. FREQUENCY INFORMATION EXTRACTOR

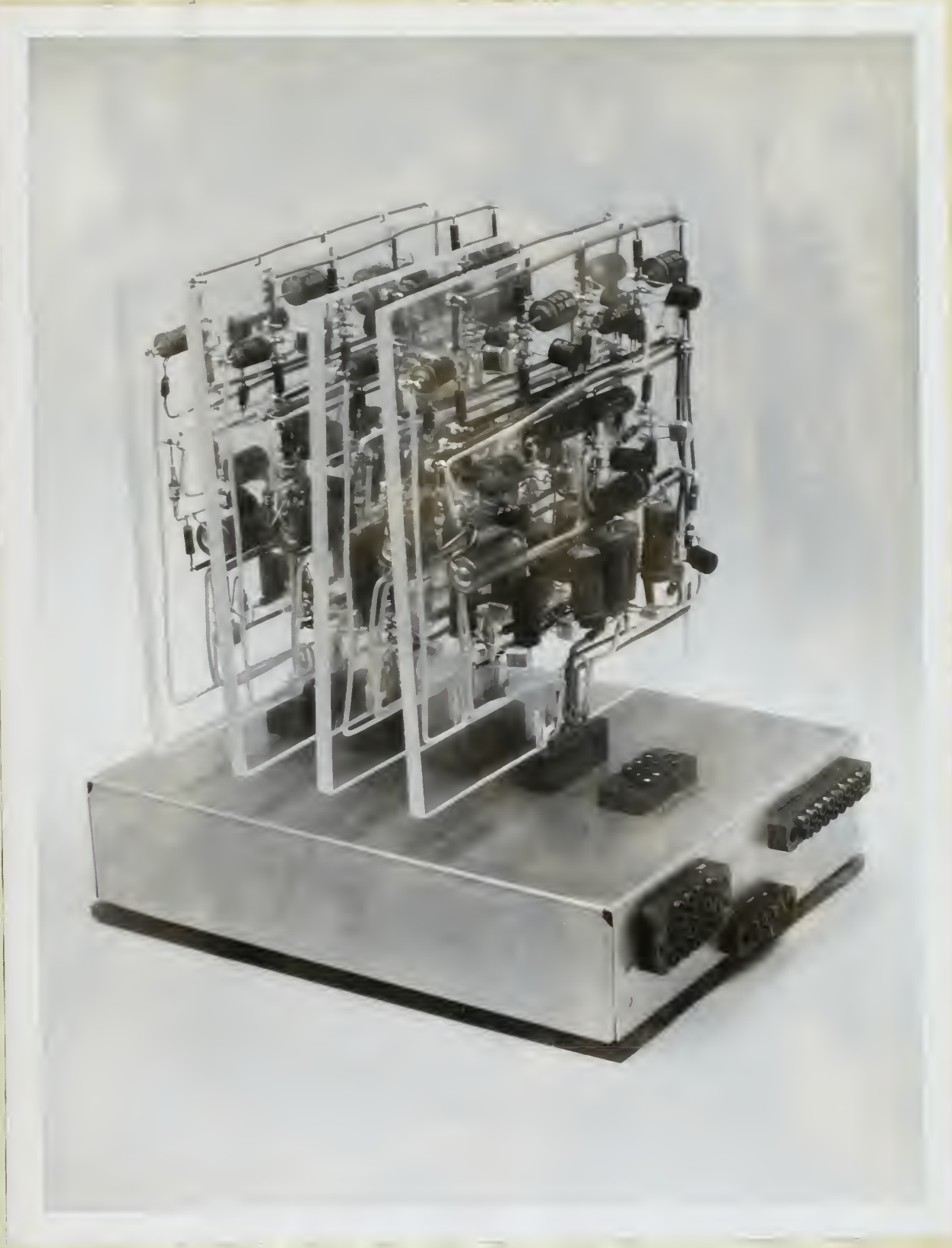


Figure 23. A photograph of the four modularized Frequency Information Extractors. The chassis contains the Pitch Extractor. The plug in devices on the front provide transistor power supply terminals, signal input-output terminals for all units, and test point terminals for access to three test points per module.

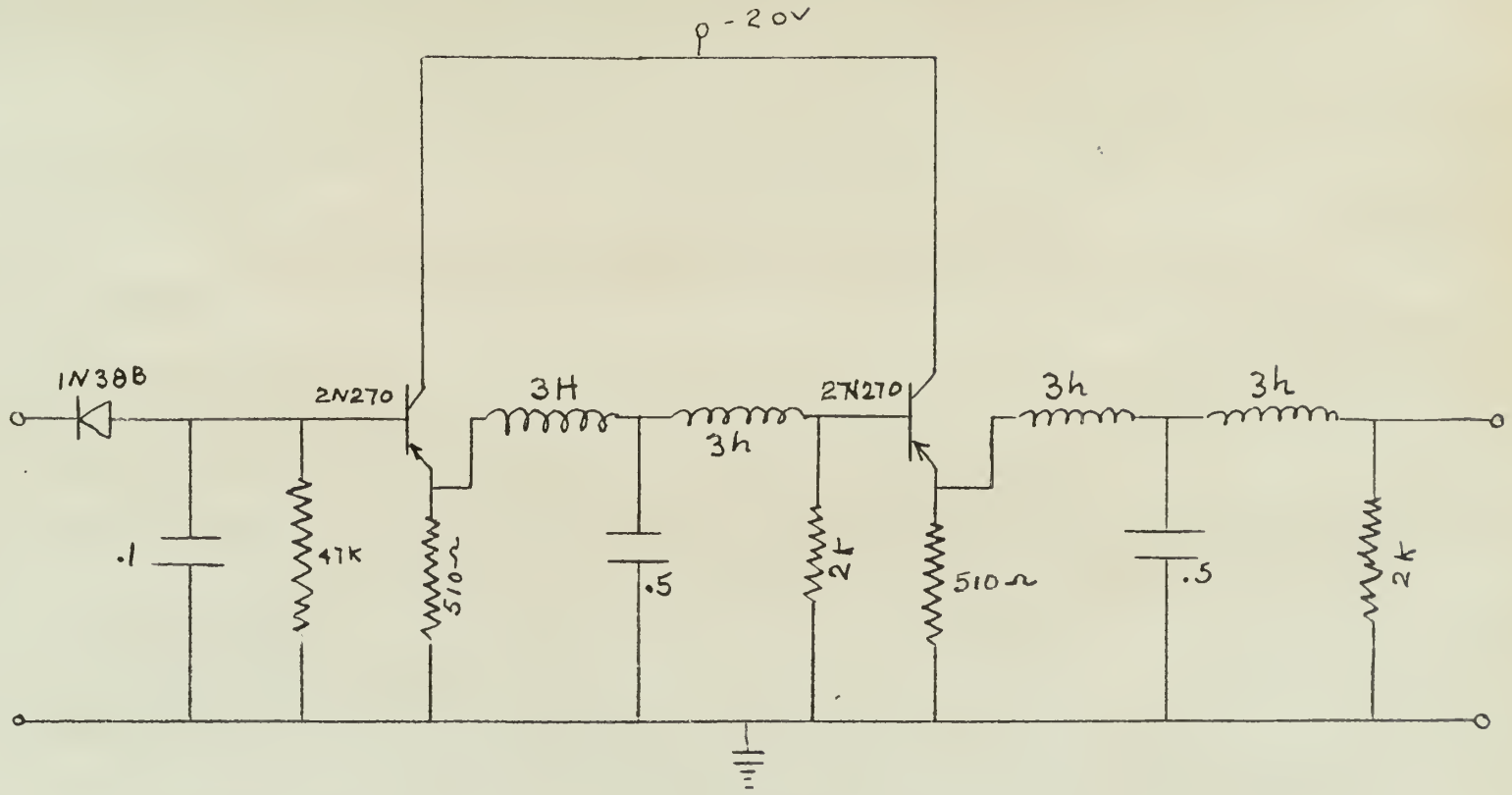


FIGURE 24. PITCH EXTRACTOR CIRCUIT

presentation to a frequency information extractor.

Fig. 16 shows the block diagram for the speech synthesis scheme. The noise generator has been previously discussed. The modulator circuitry for the synthesizer is shown in Fig. 25. Here the outputs of the various voltage variable filters are amplitude modulated by the control signals derived from the amplitude information extractors. The outputs of the modulators are resistably mixed, passed through a stage of voltage amplification and a stage of power amplification to the output speaker.

The development of a voltage variable bandpass filter for use in the audio-frequency range poses a serious problem with stringent restraints. First of all, the passband must be essentially constant for a center frequency variation of nearly 20 to 1. Secondly, the amplitude of the passed signal for a constant amplitude input wave must remain constant for the same 20 to 1 center frequency variation, namely 300 to 6000 cps.

Prior considerations as to the passbands for the filters has lead to the requirements of a 200 cps bandwidth at the half-power points for the sub-band 300 to 1500 cps; a similar 200 cps bandpass for the sub-band 1500 to 3000 cps; and a 300 cps bandpass for the sub-band 3000 to 6000. A much narrower 20 cps bandpass filter is needed in the pitch information channel.

With a view toward evaluating the stated concept of the speech analysis and synthesis bandwidth compression scheme and at the same time developing devices which could be part of a workable, non-laboratory model, it is felt that any proposed filter must be judged on a size, an economic, a weight and a simplicity criteria.

Consider first the use of standard LC filters in a T arrangement.

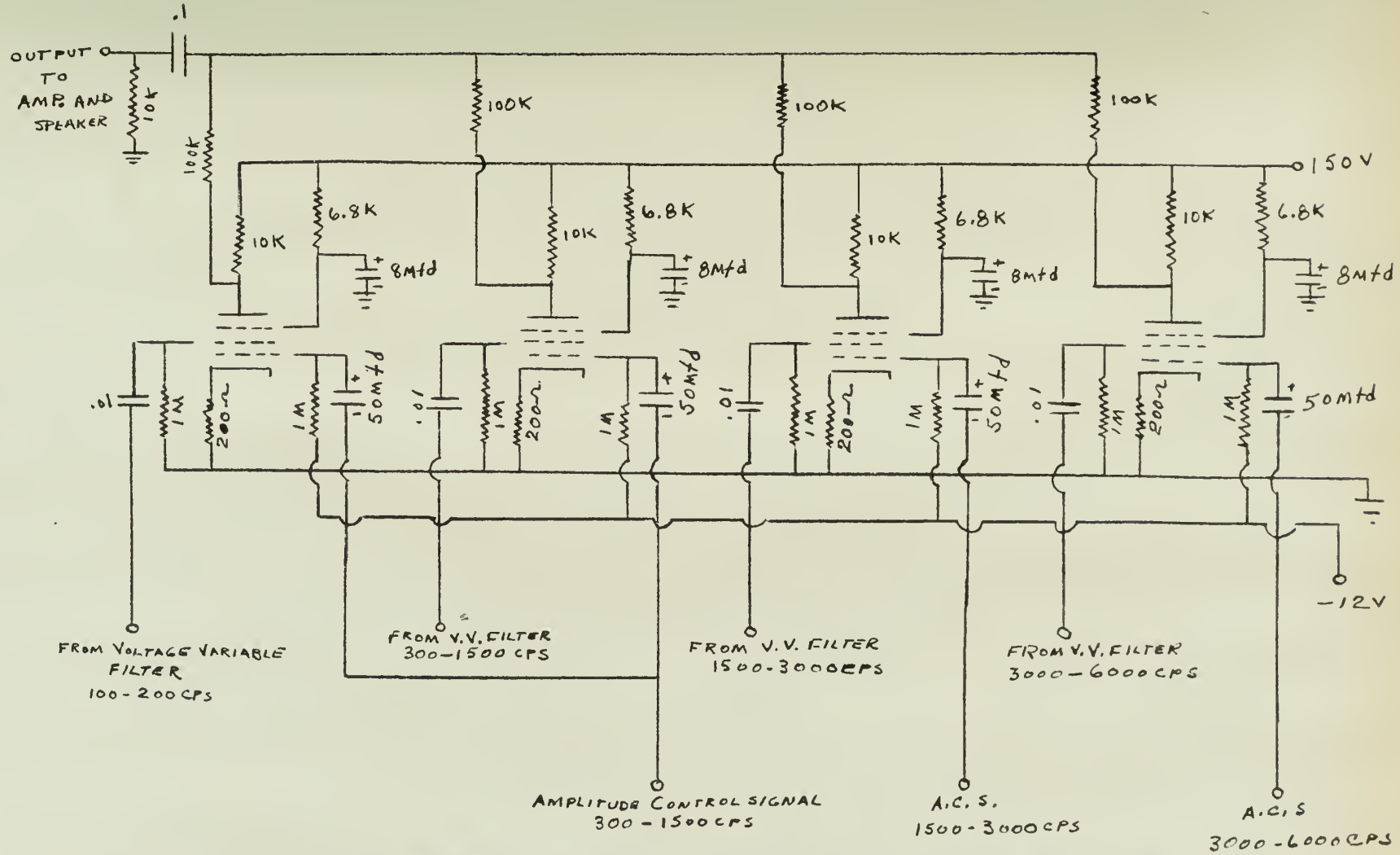


Figure 25. Circuit diagram of Modulator Unit showing inputs from voltage variable filters to be modulated by the Amplitude Control Signals, and output to amplifier and speaker. Resistive mixing of the four modulated signals is also accomplished in the Modulator unit as shown in the upper part of the diagram.

The use of this type of filter appears unprofitable from several points of view. The size of the required inductances for use in the audio region is prohibitive. The shift of the passband as a function of some control voltage requires that either the L or C components of the filter be varied continuously or in discrete steps. Variation of the L components, using Increductors,[22] to shift the passband requires sizeable auxiliary circuits. Voltage variable capacitors are commercially available at the present time, but their intrinsic capacitance and their dynamic range are as yet far too small to be of practical use in the audio region. The Q of the inductances varies with frequency, thus, the passband itself would also be a function of frequency. Also, any variation of the components to shift the passband would result in a change in characteristic impedance of the filter, so that for proper operation of the filter the terminating impedance would also have to be varied.

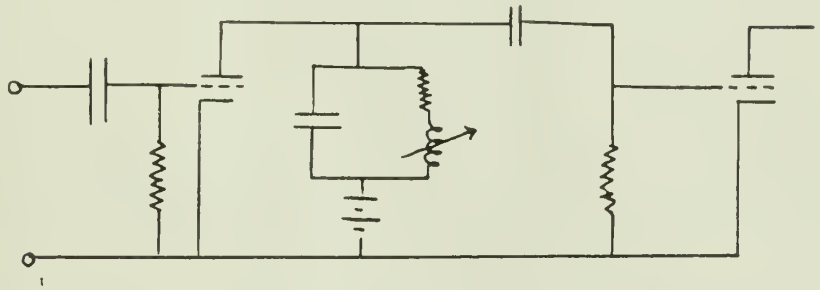
The tuned circuit provides another means by which filtering may be accomplished. Simplicity is the prime advantage of the tuned circuit filter. Here again, the high LC product required for operation in the audio region presents serious disadvantages with reference to required size and availability of suitable voltage varying components. But it is the very nature of operation of the network itself that prevents utilization of the tuned circuit in the bandwidth compression scheme. Consider the requirements for the bandpass filter. First, the bandwidth must remain constant over a wide range of frequencies. Second, the amplitude of the passed signal must not vary with frequency. The Q of the resonance curve determines the bandwidth of the filter. That is,

$$\Delta f = \frac{f}{Q}$$

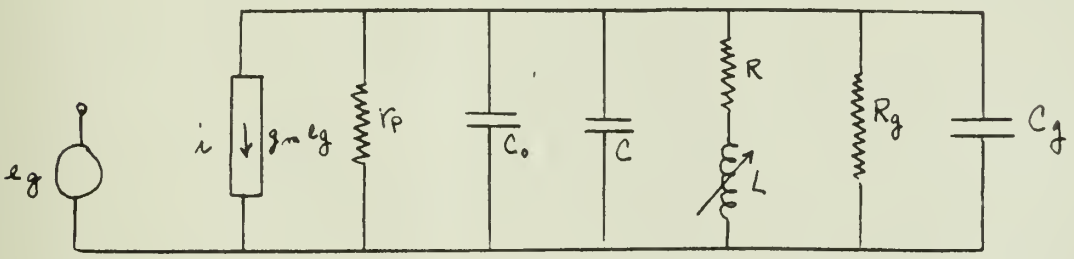
For a constant Δf bandwidth this required, for instance, in the sub-band

from 3000 to 6000 cps a Q which varies directly with frequency. At 3000 cps for a bandwidth of 300 cps the required Q is 10; for 6000 cps, a Q of 20. As the frequency increases, so must Q .

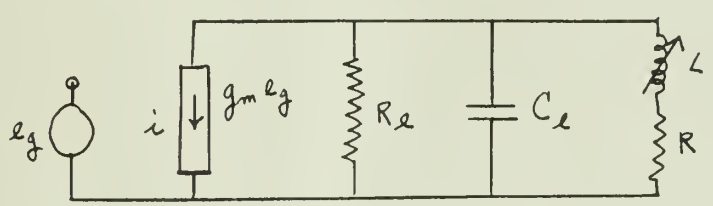
The following circuit is a simple tuned circuit bandpass amplifier where the passband is shifted by means of Increductors,



The equivalent circuit of this tuned amplifier is:



This circuit may be redrawn as follows.



$$C_e = C + C_o + C_g$$

$$\frac{1}{R_e} = \frac{1}{r_p} + \frac{1}{R_g}$$

The resonant frequency of this circuit:

$$f = \frac{1}{2\pi\sqrt{LC_e}} \quad \text{OR} \quad \omega_o^2 = \frac{1}{LC_e}$$

The Q is equal to:

$$Q = \frac{1}{\frac{\omega_o L}{R_e} + \frac{R}{\omega_o L}} = \frac{\omega_o L}{\frac{\omega_o^2 L^2}{R_e} + R}$$

Substituting for L, $L = \frac{1}{\omega_o^2 C}$ $\quad Q = \frac{\frac{\omega_o}{C}}{\frac{1}{C^2 R_L} + \omega_o^2 R}$

Thus, it is seen that as the passband is shifted by varying L, the Q of the circuit varies inversely with frequency. That is, as the frequency increases, Q decreases. This is exactly opposite to the required performance of the filter. Therefore, the use of a tuned filter is not possible in this case with the stated specifications. Also consider, if the amplitude of the passed signal is to remain constant throughout the sub-band, the impedance of the circuit as seen by the current generator must remain constant.

Neglecting the shunt capacitances, the load for the current generator is $\frac{1}{Z_L} = \frac{1}{R_L} + \frac{1}{Z_{TANK}}$

$$\left| Z_{TANK} \right| = \left[\frac{R^2 + \omega^2 L^2}{(1 - \omega^2 LC)^2 + \omega^2 C^2 R^2} \right]^{1/2}$$

Now, as L is varied to shift the passband, as shown below, it can now be seen that as the passband is shifted and the center frequency increases, the magnitude of Z_t decreases. R is a constant so that the magnitude of the load for the current generator varies with the magnitude of the tank impedance. Thus, the output amplitude varies with the center frequency and the second requirement of the filter is not met.

Substituting $L = \frac{1}{\omega_o^2 C}$

$$\left| Z_{TANK} \right| = \left[\frac{R^2 + \frac{1}{\omega_o^2 C^2}}{\omega_o^2 C^2 R} \right]^{1/2}$$

Investigations have been made using the tuned circuit as a bandpass

filter in the audio region and accepting the resulting filter limitations[16,17]

A very excellent continuous filter has been designed and built by Fant.[23] Filtering is accomplished by a series of heterodyning actions using fixed filters. The heterodyne filter provides a constant bandwidth which is essentially independent of audio frequency. The bandwidth is also easily modified. Fant's filter provided cutoffs with a maximal slope of 1 db per cps and an ultimate of over 60 db of attenuation. The operation of the heterodyne filter is shown in Fig. 26. The heterodyne filter as designed by Fant was to operate in the 45 to 4000 cps region. Therefore, in Fig. 26, the input signal is passed through a low-pass filter, removing components of the spectra above 4000 cps. The ultimate object of the filter is to pass a band of frequencies Δf from F_L to F_H as shown in Fig. 26. As shown in Fig. 26b, the input signal from the low-pass filter of Fig. 26a is heterodyned with a frequency $f_1 = F_1 + F_L$ where F_1 is the fixed cutoff frequency of the low-pass filter shown in Fig. 26b, and F_L is the desired lower limit of the ultimate passband. The resulting signal has the upper sideband removed by the low-pass filter of Fig. 26b. The lower sideband is passed through the bandpass filter, Fig. 26c. The signal from the bandpass filter is then heterodyned with f_2 whose placement along with the cutoff frequency F_2 of the low-pass filter of Fig. 26c determines the desired upper frequency F_H of the ultimate passband. The remaining band of frequencies is then heterodyned with f_3 to place it in its proper position in the frequency spectra.

The high performance and versatility of the heterodyne filter has much to offer. Unfortunately, the complexity and size of the circuitry, Fant's filter was an eight rack device, precludes any reasonable use in

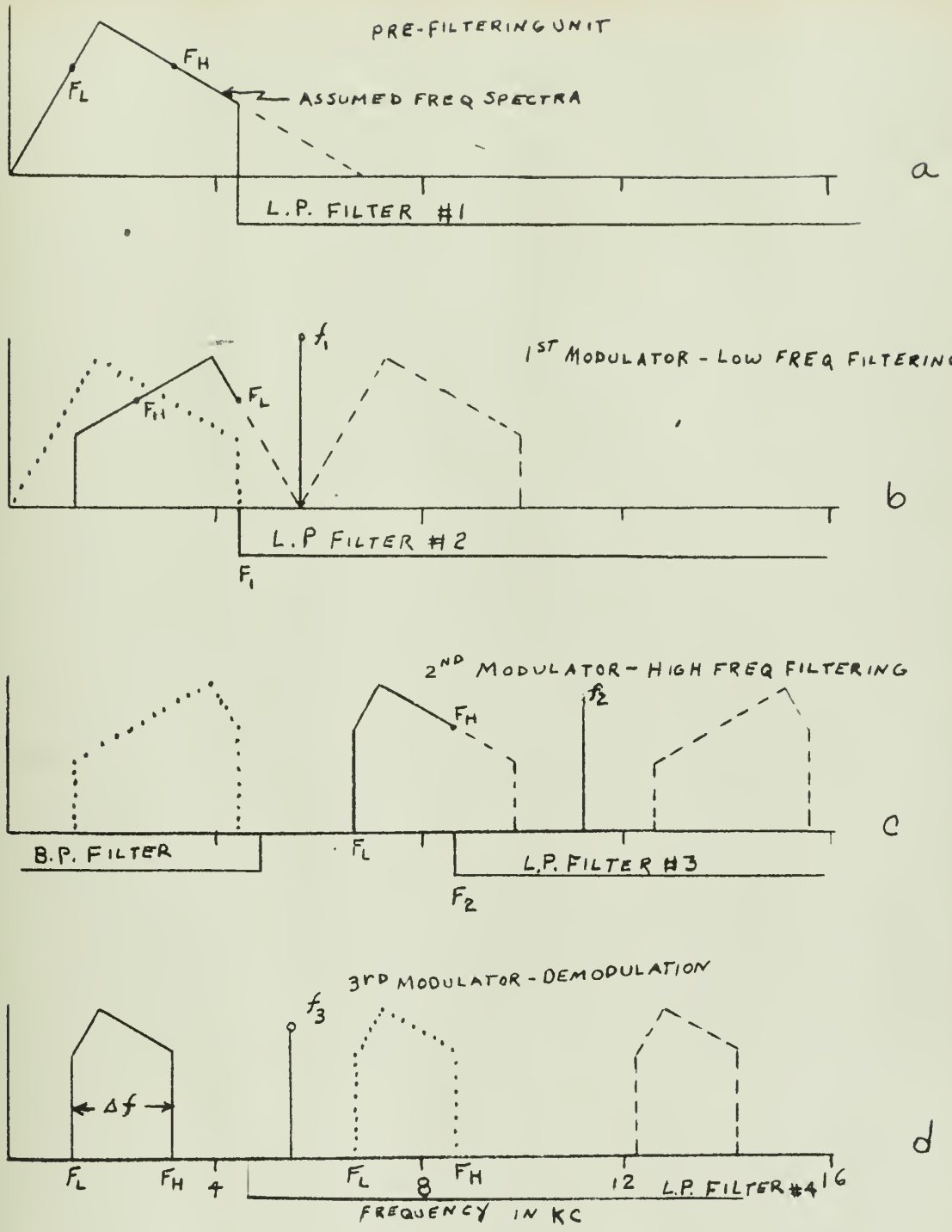


FIGURE 26. CONTINUOUSLY VARIABLE BANDPASS FILTER SCHEME. CONVERSIONS SYMBOLIZED BY AN ASSUMED FREQUENCY SPECTRA OF INPUT SIGNAL.

• • • INPUT SIGNAL BAND

_ _ _ FREQ. BAND ATTENUATED

___ OUTPUT SIGNAL BAND

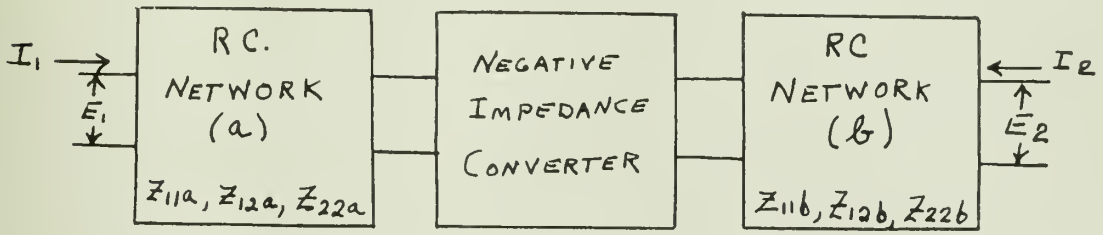
a workable model.

The use of crystal filters and heterodyning techniques holds high promise as an efficient means to accomplish the required filtering. High stability, variable frequency oscillators are the prime requirement of this approach. The concept here is to mix the band of frequencies in the audio region to be filtered with some variable, high-frequency, carrier, pass the lower or upper sideband through a fixed crystal filter, and then heterodyne the passed band back to the audio region. The passband shift would be accomplished by moving the sideband relative to the fixed crystal filter by varying the initial high-frequency carrier. This system has the attribute of constant bandwidth and constant amplitude output for a constant amplitude set of input audio frequencies. Variations in the desired passband may be accomplished by two means. Crystals having the same resonant frequency but different Q's may be picked; or a crystal having a higher resonate frequency but a given Q may be chosen. The bandwidth being determined by $\Delta f = \frac{f_o}{Q}$. For example, a crystal having a resonant frequency of 1 megacycle and a Q of 20,000 provides a bandpass of 50 cps, while a crystal having the same resonant frequency but a Q of 10,000, has a bandpass of 100 cps. Similarly, a 2 megacycle crystal with a Q of 20,000 has a bandpass of 100 cps and a 2 megacycle crystal of a Q of 10,000 has a 200 cps bandwidth.

This particular approach was not followed in the investigation because of a desire to find an equally efficient method of filtering in which the filtering would be done in the audio region. Thus, the problems of high stability oscillators, heterodyning, and a larger volume of circuitry could be avoided.

RC active filters for high-pass, low-pass and bandpass filters, have

provided the basis in recent years for another type of electronically controlled audio filter[22]. The RC active filter has the ability to provide characteristics corresponding to those of the usual types of RLC passive filters. In this device, a negative impedance converter is used in addition to passive RC elements. The sum of the capacitors in the circuit is equal to the sum of the reactances in the corresponding RLC filter. The normal inband loss associated with RC passive filters are greatly reduced by the active element. The block diagram for a RC active filter is shown below.



The negative impedance converter is an active four-terminal, four-pole, which presents at the input terminal pair the negative of the impedance connected to the output terminal pair.[24] The transfer impedance for a lumped element filter may be written $Z_T(s) = \frac{N(s)}{D(s)}$ which for the RC active filter is

$$\left.\frac{E_2}{I_1}\right]_{I_2=0} = Z_{21} = \frac{Z_{12a}\, Z_{12b}}{Z_{22a} - Z_{11b}}$$

where the negative sign before Z_{11b} is provided by the negative impedance converter.

The design of the circuit is basically simple. The zeros of D(s) are chosen at the desired natural frequencies of the completed structure. From this the driving point impedance for the structures a and b are calculated. The structure form is selected to provide zeros of transmission

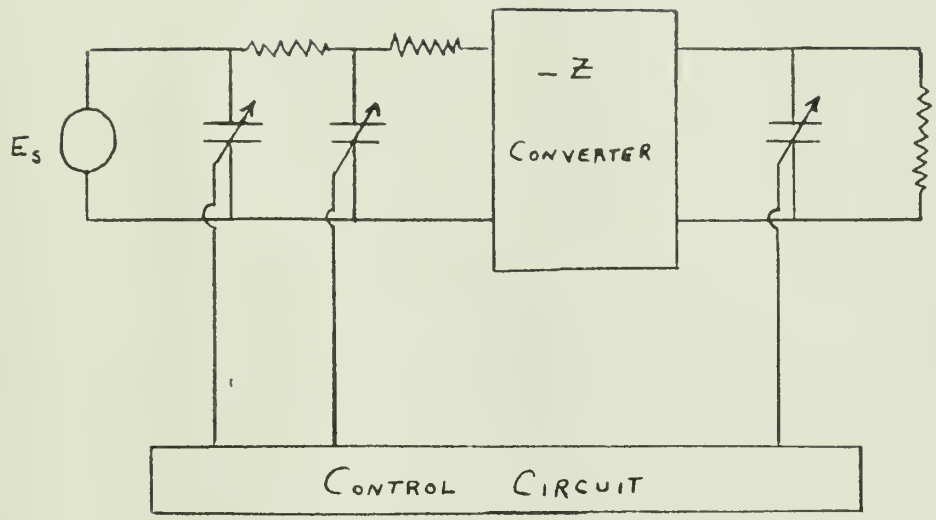
at the required frequencies, these are the zeros of N(s).

Based on the work of Linvill, Dolansky developed voltage variable high-pass and low-pass filters which, when arranged in series, provide a voltage variable bandpass filter. The simplified diagrams of the low and high-pass filters are shown in Fig. 27. Using the Miller effect, to provide the voltage variable capacitance, Dolansky's circuit required a three-tube circuit per variable capacitance in the control stage. The variable inductances are saturable inductors whose inductance depends upon the degree of core saturation.

The audio filter as developed by Dolansky provided a cutoff slope of 17 db per octave. The use of Increductors in the circuits, although providing the variation required, leads to undesirable effects. Hysteresis causes the inductance to vary about 10 percent for the same control current. The bandpass varies with frequency because of the Q variation in the inductance. The circuitry is sizable.

It is thus felt this time that there are better and simpler circuits to provide a variable filter.

The approach taken and the voltage variable filter that was designed and built for the investigation may at first appear to be awkward and to be the hard way of doing things. But, the system was developed with the future state of the art in mind. It is felt that within two or three years, a voltage variable capacitance will be produced, having the required intrinsic capacitance and dynamic range, that will make the design system a highly efficient but simple method for variable filtering in the audio region. It is believed that the superiority of the system that can be attained with the use of proper voltage variable capacitors more than offsets the circuit complexity needed at present to implement the concept



VARIABLE LOW PASS ACTIVE FILTER

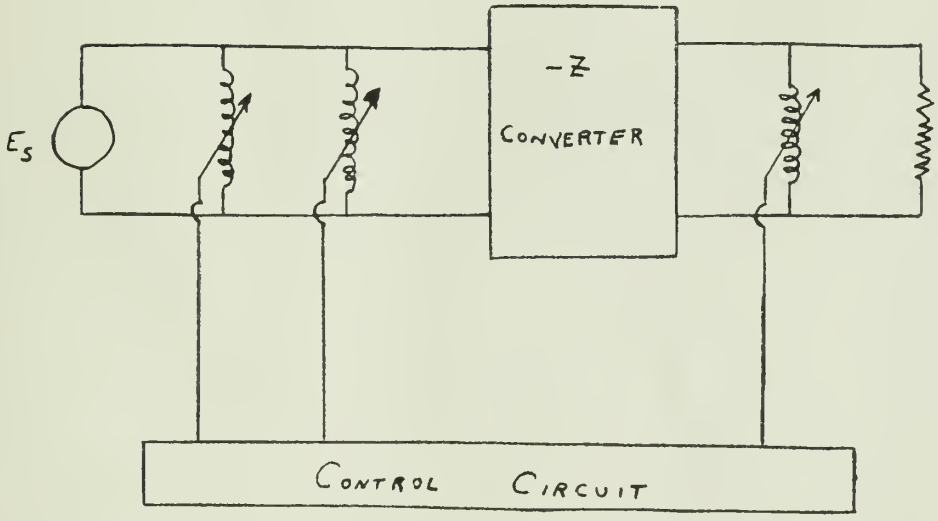
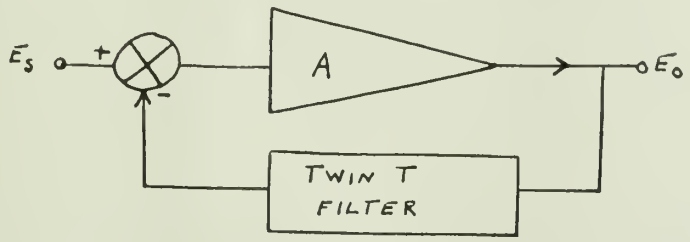


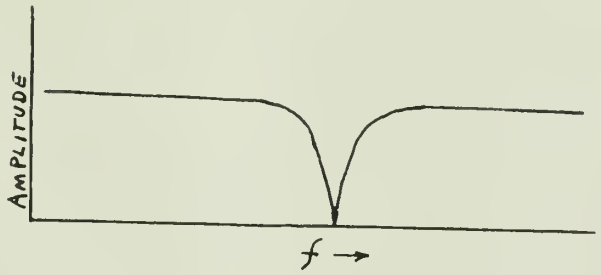
FIG. 27 VARIABLE HIGH PASS ACTIVE FILTER

with currently available components.

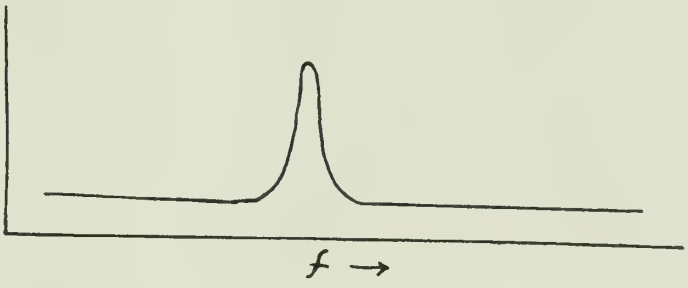
The filter consists of a Twin T rejection filter in the negative feedback path of an amplifier. The gain of the amplifier is reduced by the negative feedback at all frequencies except the rejection frequency of the Twin T filter. A block diagram for the system is shown below.



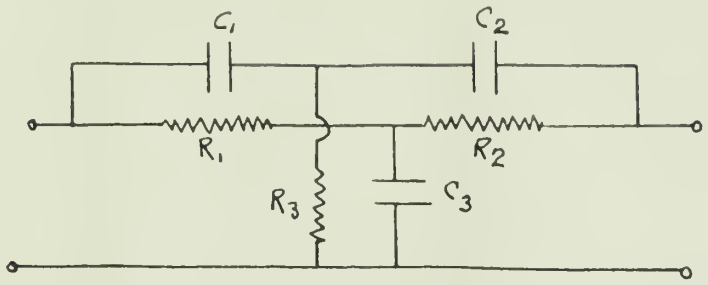
The transfer curve for the Twin T Filter is



As is seen, the Twin T passes all frequencies except those in the notch. Thus, there is negative feedback to the amplifier at all frequencies except the rejection frequency. The resulting characteristic for the system is then



The Twin T circuit consists of three resistances and three capacitances as shown below.



The problem of making the Twin T filter voltage tunable, varying in accordance with some control voltage poses an interesting problem. In order to shift the rejection frequency of the Twin T and thus shift the passband of the filter, either all of the resistive elements or all of the capacitances must be varied together. Voltage variable resistors are available, but they are non-linear with voltage and the problem of maintaining a match between resistors is extremely difficult. Another interesting scheme considered was to use a photocell as a variable resistance. The resistance is varied by changing the light intensity incident upon the photocell. A neon tube was considered as a possible light source. Experiments showed that the light intensity eminating from the neon tube to be non-linear with voltage except in narrow regions. The possibility of using a magic eye tube and intensity modulating the electron flow was considered. This approach appears to have some merit, but was not fully investigated as the basic plan to use photocells as a variable resistance proved too difficult to implement. It is very difficult to match photocells, both dynamically and statically, to give the same resistance for the same light intensity.

Voltage variable capacitors, Vericaps currently available, as has

been said, do not possess the proper parameter size and range for use in this frequency region. Currently available Vericaps have a maximum capacitance of the order of 300 to 350 micromicrofarads. It is felt that, when the capacitance of available Vericaps is of the order of 1000 micromicrofarads or larger, they may be used practically as voltage variable components in the Twin T filter.

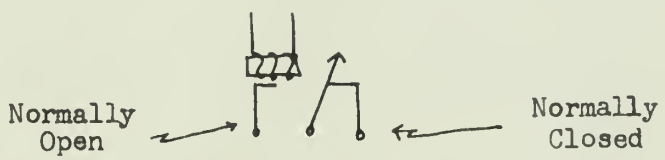
The junction capacitance of a semiconductor diode, as is well-known, is voltage variable. As the back bias to a p-n diode is varied, the barrier width changes and thus, its capacitance. Experiments conducted on a 1N1084 silicon diode showed a 168:1 variation in capacitance for a back bias variation of approximately 50 volts. Unfortunately the non-linearity of the capacitance and the difficulty in matching diodes precluded their use in the circuit.

The solution of the problem lead to a circuit which, aside from providing a voltage tune filter, is unique in itself. The circuit is a marriage of transistors, tubes, and relays. Due to the great difficulty in obtaining, at this time, continuously voltage variable components, and thus enjoying a continuously variable filter, it was decided to vary the components in discrete steps and thus obtain a discrete rather than continuous filter. It must be emphasized that the restriction to discreteness will be removed with the expected advent of Vericaps possessing the proper parameter size.

The method of discretely shifting the bandpass is to change the values of all three resistive components of the Twin T together by the use of relays. The control voltage or shifting the passband of the filter is fed to a transistor relay control network. As the control voltage rises, a series of relays are closed. Each relay closing at a given control

voltage value as determined by the relay control network. As each relay closes, the three resistances of the Twin T are changed. The rejection frequency of the Twin T is changed and thus the passband of the filter is shifted.

Consider first the relay control network as shown in Fig. 28. The function of this circuit is in serial fashion to cause a set of relays to be picked up. The control voltage varies from minus 20 volts to ground potential. The control sequence is as follows: When the control voltage is at minus 20 volts, all of the 2N441 transistors are cut off causing all of the relays to be open. When the control voltage rises to a less negative potential which is equal to the negative potential of the emitter of the 2N214 transistor associated with number one relay, the 2N214 moves from cutoff to an operating position. This action causes a large current to flow in the relay pickup coil, due to the current and power gain of the 2N270 and 2N441 circuitry. Relay one is thus closed. The emitters of the various 2N214 transistors are set from left to right at progressively more positive potentials, the individual values being at the desired control voltage values for the closing of the relays. Thus, as the control voltage rises from minus 20 volts to ground, the relays close in serial fashion from left to right at a predetermined control voltage value. The relays used were IBM type 104753. These relays are four-terminal set devices enabling the relay to control four separate circuits, four elements independently as it opens and closes. Symbolically, one of the four terminal sets of the relay is shown below.



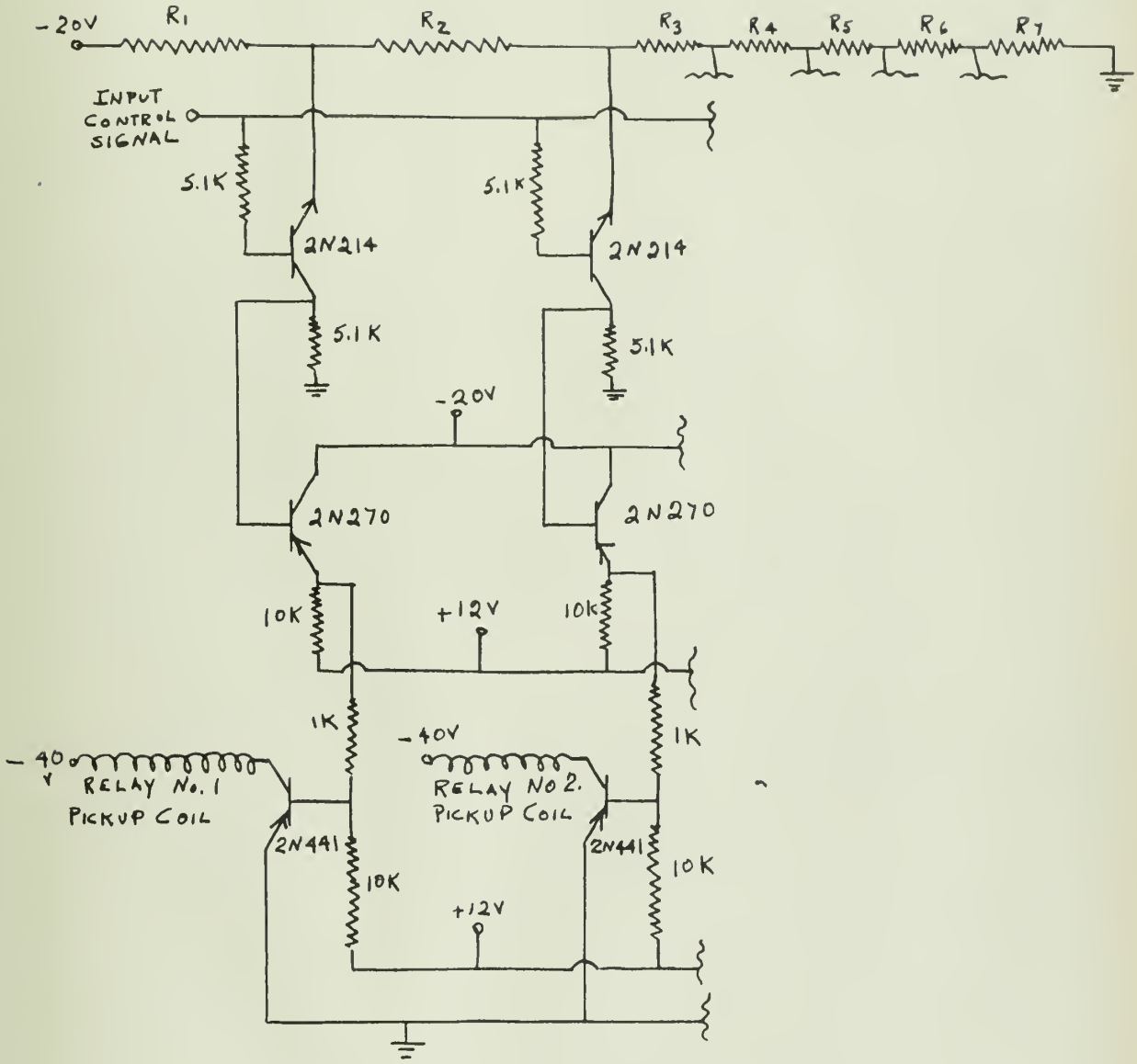


Figure 28. Relay Control Network Showing First Two Stages

With the relay non-energized, the control arm rests in the normally closed position. When energized, the control arm shifts to contact the normally closed position.

It must be realized that the pickup time of the relay is finite, being of the order of approximately 3 milliseconds. It is felt that this pickup time is well within the demands of the system, as the control voltages will vary at approximately a 20 cps rate.

Variations in the circuitry of the relay terminal sets allow three different methods of changing the resistive components of the Twin T. The resistances may be changed by adding in series discrete resistances, by paralleling resistances, or by causing the relays to place in or take out individual resistances. The parallel method is shown in Fig. 29. This method is not recommended as the resistance values associated with each step tend to become very large and thus have a higher level of thermal noise.

The series method is illustrated in Fig. 30 and the individual method in Fig. 31. Both the series approach and the individual component approach were utilized in the system in order to experimentally determine the relative merits of the two. The series approach has the advantage of wiring simplicity in regards to connections between the resistors of the matrix and the therminal sets of the relays. Resistance values per terminal set are naturally lower in value. The individual component approach was found to be the best system. In the individual system, each position of the bandpass may be set up and tuned without regard for any of the other setups for other bandpass positions. In the series approach, if for a given control voltage a different frequency for any one of the steps is desired, the entire resistance matrix associated with each arm of the matrix must be changed.

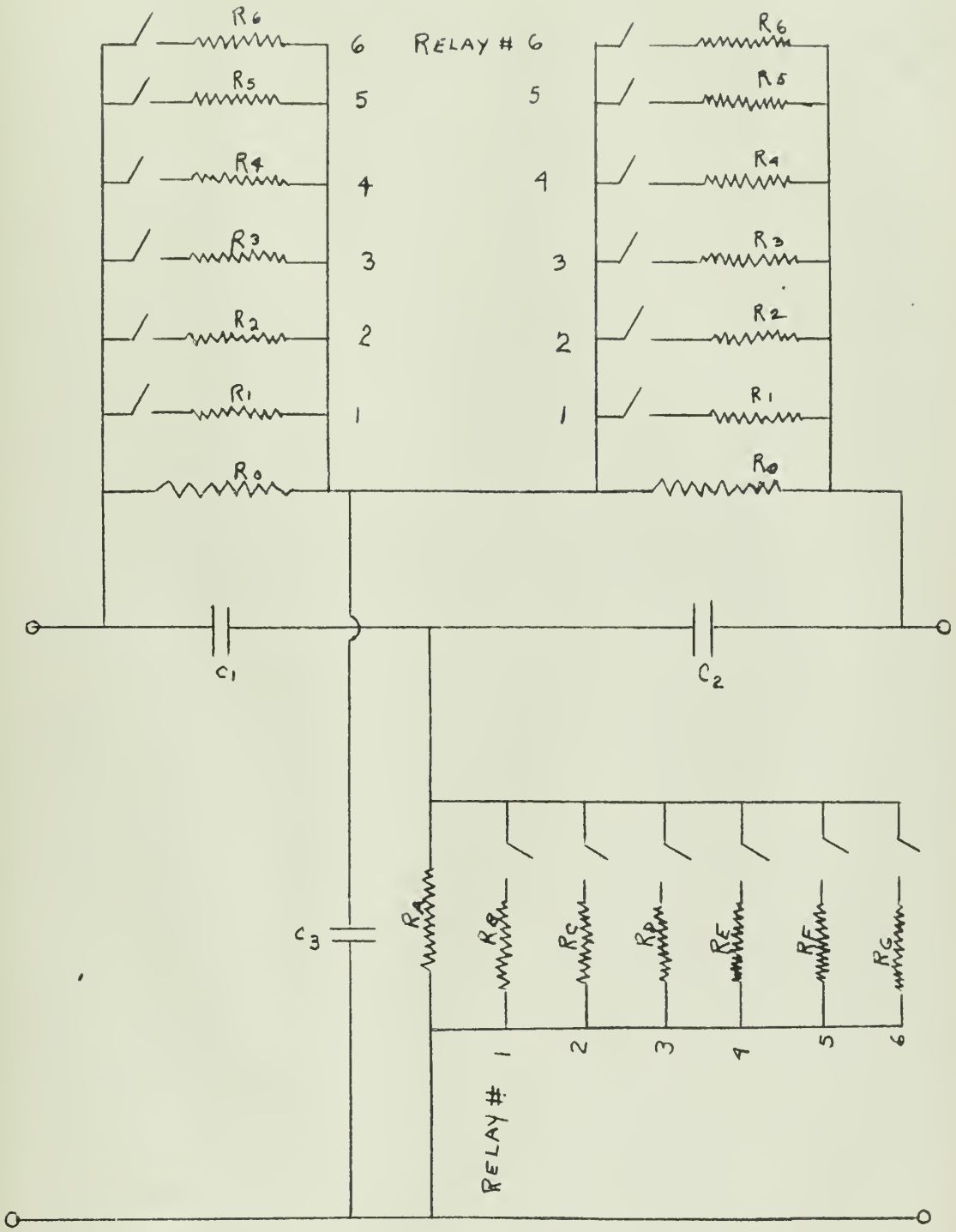


FIG. 29 PARALLEL METHOD OF CHANGING RESISTIVE ELEMENTS OF TWIN T

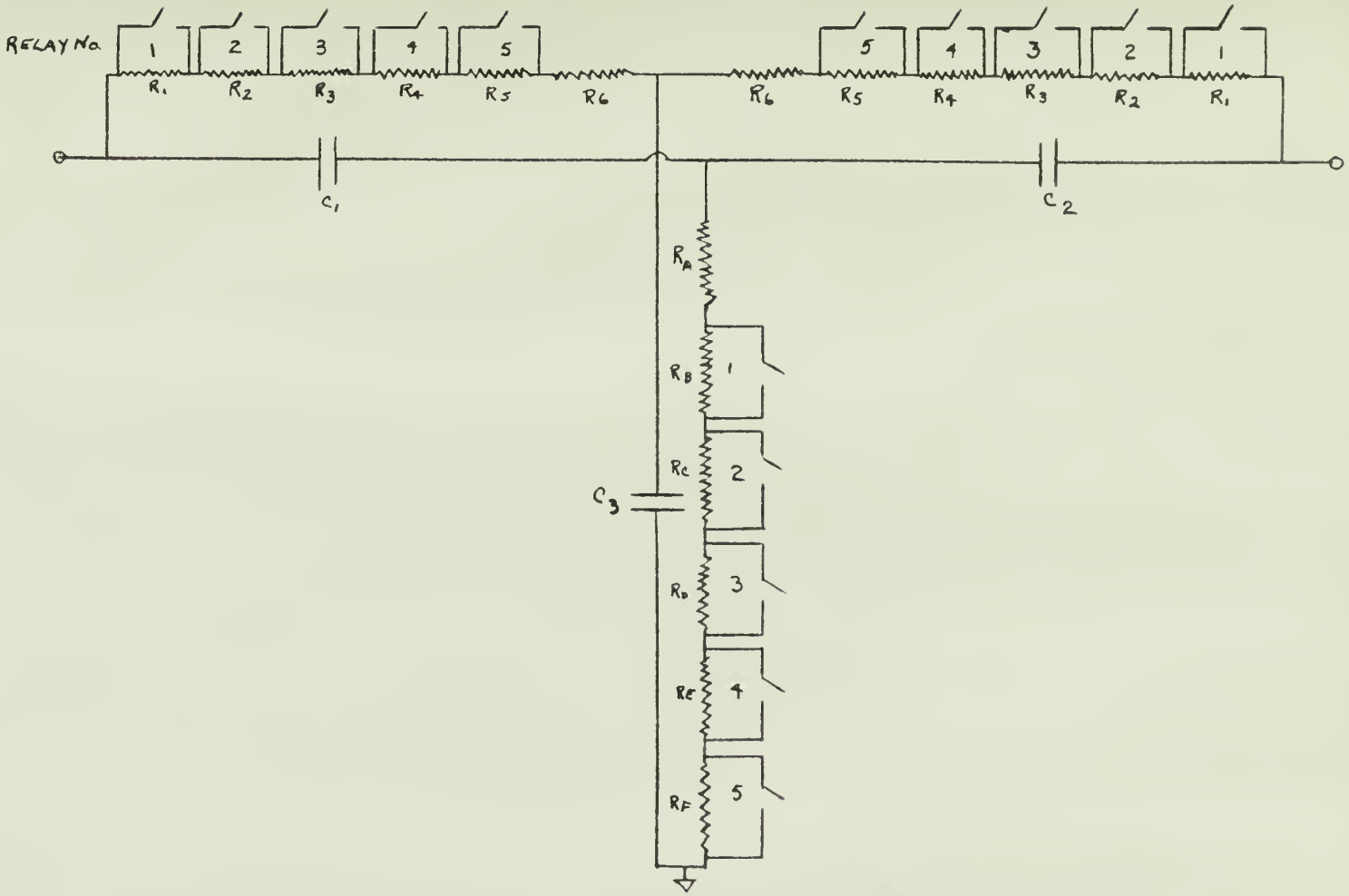


FIG 30. SERIES METHOD OF CHANGING RESISTIVE ELEMENTS OF TWIN T.

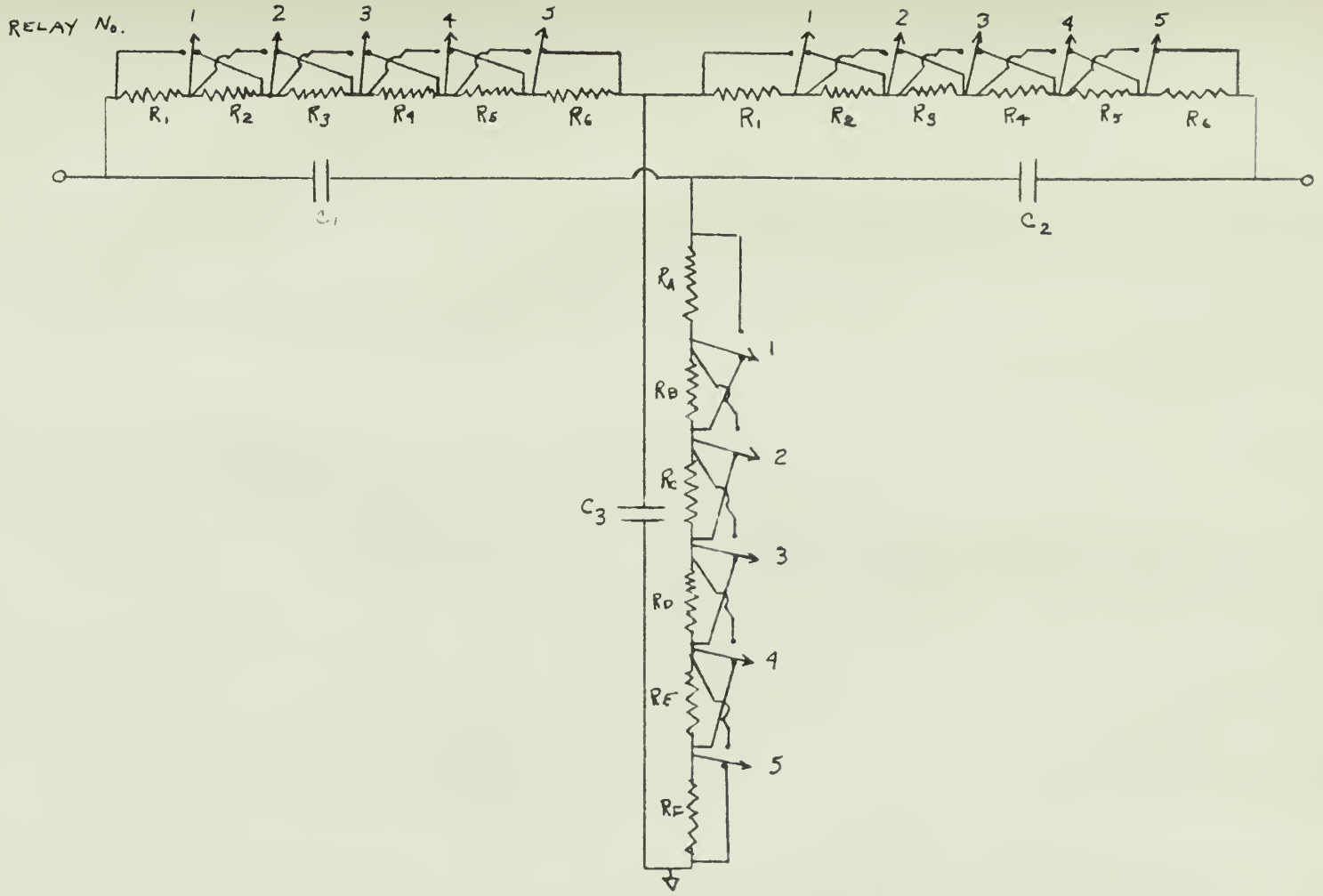
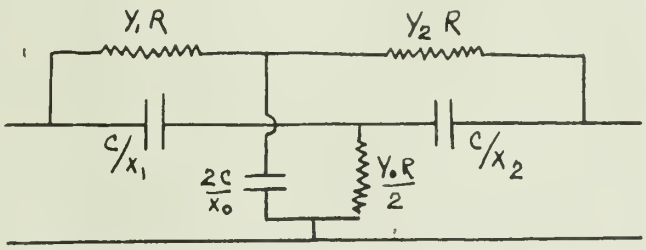


FIG 31. INDIVIDUAL METHOD OF CHANGING RESISTIVE ELEMENTS OF TWIN T.

Inasmuch as the value of the effective resistance in any arm for a given bandpass position is extremely critical, any variation in the desired center frequency entails an inordinate amount of labor.

It was found that the actual design and construction of the Twin T filter tends, as various references in the literature subtilely imply, to be more of an art than a science. On this basis, the inclusion of some of the empirical procedures determined in the construction of the filter is deemed to be warranted in this paper.

The basic theory of the Twin T will first be investigated.[25] In generalized form, the parameters of the Twin T, as shown below, must conform



to the following relationship for any given rejection frequency.

$$(1) \qquad X_0 Y_0 = \frac{2 X_1 X_2}{X_1 + X_2} \cdot \frac{2 Y_1 Y_2}{Y_1 + Y_2}$$

The rejection frequency is then given by

$$(2) \qquad f = \sqrt{\frac{X_0 (X_1 + X_2)}{2 Y_1 Y_2}} \quad \frac{1}{2\pi C R}$$

Various degrees of sharpness in the rejection characteristic at any frequency may be obtained by proper manipulation of the above equations. A measure of rejection sharpness is given by equation (3).

$$(3) \qquad A = \left[\frac{X_1 + X_2}{4 Y_2} + \frac{X_1}{2 Y_0} \right] \sqrt{\frac{2 Y_1 Y_2}{X_0 (X_1 + X_2)}}$$

Sharpness of rejection is indicated by lower values of A. For a symmetric

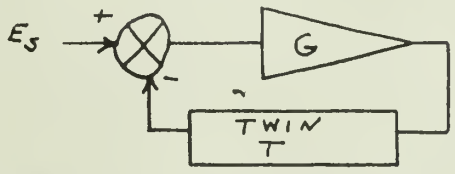
configuration, $X_1 = X_2$, $Y_1 = Y_2$, the smallest A obtainable is A = 1 and occurs when $X_o = Y_o = 1$.

By going to an unsymmetric network, smaller of A may be obtained. A convenient design for the unsymmetric T is

$$X_1 = Y_1 = 1 \qquad X_o = Y_o = \frac{2k}{1+k}$$

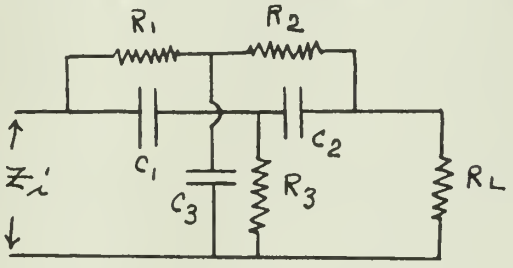
$$X_2 = Y_2 = k \qquad A = \frac{1+k}{2k}$$

The network that is most usually encountered is the symmetric network, for which $X_1 = X_2 = X_o = Y_1 = Y_2 = Y_o = 1$. Tucker[26] has shown that when a Twin T which is symmetric with A = 1 is included in the negative feedback path of an amplifier with gain equal to G, as shown below,



the Q of the system as a passband filter is $Q = \frac{G}{4}$.

Scott[27] has shown that the input impedance of the Twin T is



(4) For $R_L \geq 3(R_1+R_2)$ $\qquad Z_i = V\left[1 - j\frac{Q}{M}\right]PR$

$$V = \frac{q - \frac{f}{K}}{1 - j\left(\frac{P+Q}{M}\right)} \qquad P = \frac{abq}{a+bq^2} \qquad Q = \frac{\alpha\beta q}{\alpha+\beta q^2} \qquad R = R_1 + R_2 \qquad K = \frac{f}{f_o}$$

$$M = \frac{K}{1-K^2} \qquad a = \frac{R_1}{R_1+R_2} \qquad \alpha = \frac{R_2}{R_1+R_2} \qquad q = \frac{X_1+X_2}{R_1+R_2} \qquad X_1 = \frac{1}{\omega C_1} \qquad b = \frac{X_1}{X_1+X_2}$$

$$\beta = \frac{X_2}{X_1+X_2} \qquad f_o = (2)$$

Four voltage variable bandpass filters were constructed during the investigation. Filter #1 covered the audio spectrum from 100 to 200 cps; Filter #2 from 300 to 1500 cps; Filter #3 from 1500 to 3000 cps; and Filter #4 from 3000 to 6000 cps. Table 1 shows the center frequencies and bandpass characteristics for the various filters.

Table 1

Filter #1 Bandwidth 20 cps	Center Freq. cps	Filter #2 Bandwidth 200 cps	Center Freq.
	100		300
	110		500
	125		700
	145		900
	170		1100
	200		1300
			1500

Filter #3 Bandwidth 200 cps	Center Freq.	Filter #4 Bandwidth 300 cps	Center Freq.
	1700		3200
	1900		3600
	2100		4000
	2300		4400
	2500		4800
	2700		5200
	2900		5600

For discussion purposes, the development of Filter #4 will be described.

The circuitry for Filter #4 is shown below in Fig. 32a. The construction of the Twin T and its associated relays are shown in Fig. 32b. The circuit is seen to consist of a cathode follower, the Twin T matrix with its associated relay control network, and a stage of amplification. If the Twin T is set for a rejection frequency of say 4000 cps, then there is no negative feedback to the grid of the cathode follower. A signal of 4000 cps is then permitted to exist at the cathode of the cathode follower

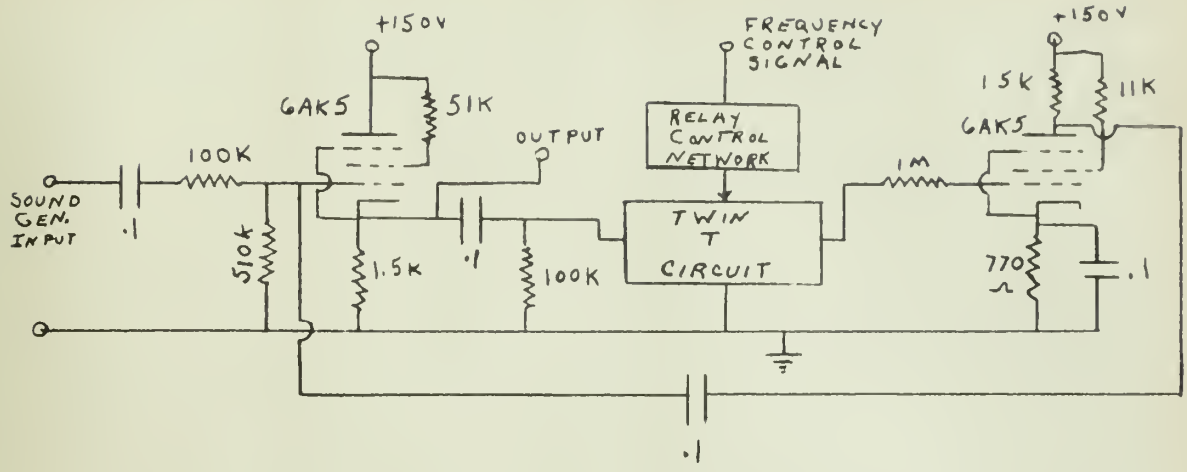


FIG. 32a. VOLTAGE VARIABLE FILTER #4, 3000 – 6000 CPS

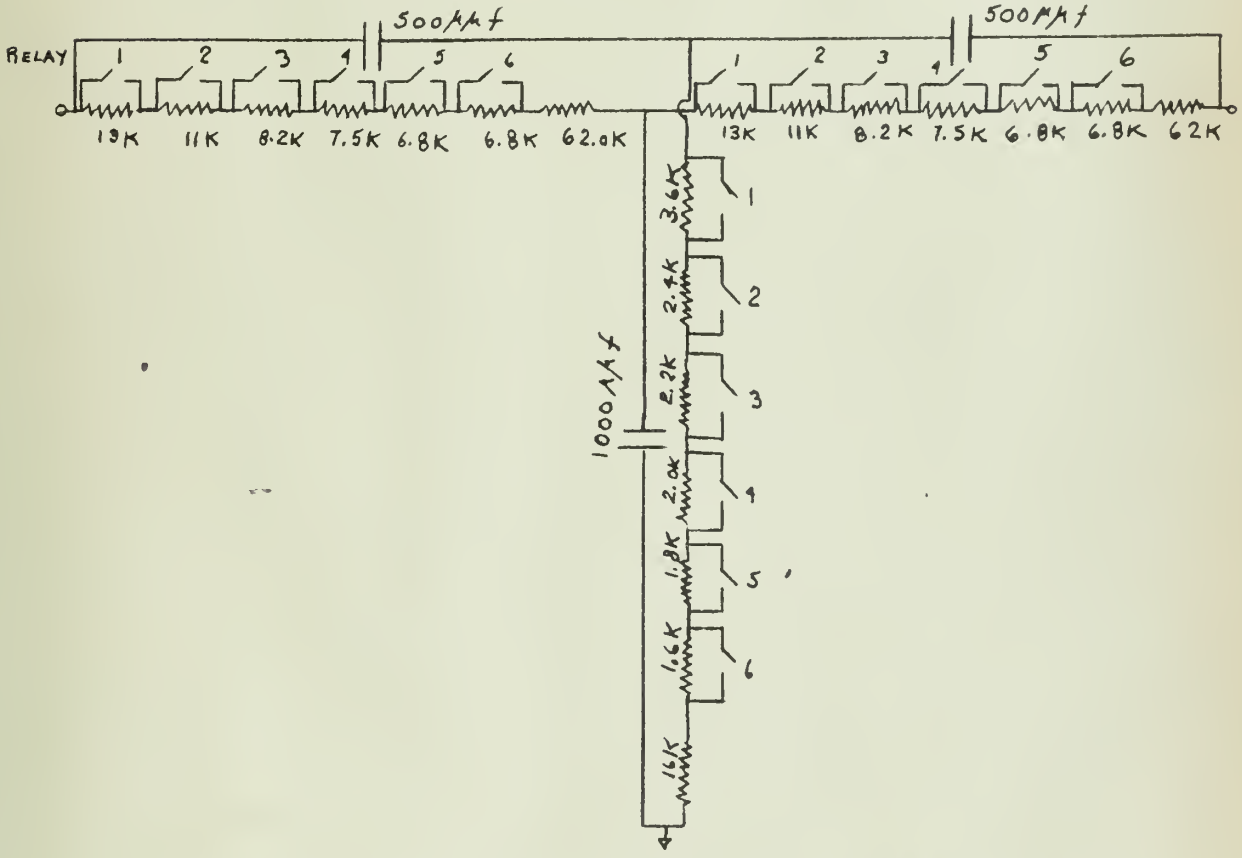


FIG 32b. TWIN T. CIRCUIT FOR VOLTAGE VARIABLE FILTER #4.

when the incoming signal is 4000 cps. If the input frequency is changed, the Twin T passes this frequency and there is a negative voltage feedback to the grid of the cathode follower. The output of the system is thus reduced.

Originally, the position of the cathode follower and amplifier were interchanged. It was found that better impedance conditions and less hum were encountered in the configuration shown. The Twin T requires a load impedance at least three times as great as the sum of its series resistances.

The input to the system is voice characterized, band limited sound, band limited here to a region of 3000 to 6000 cps, obtained from the #4 sound generator. The action of the filter is to select from this 3000 to 6000 cps sub-band a smaller band 300 cps in width, the particular smaller band chosen being determined by the control voltage. There are seven small bands associated with this filter. The center frequencies of the bands being given in Table 1. When all the relays are open, the small band selected is the lowest frequency band in the sub-band. When the DC level of the control signal corresponds to a frequency of 3400 cps, relay 1 closes and the smaller band selected is centered at 3600 cps. As the DC level of the control signal varies but corresponds to any frequency from 3400 to 3800 cps, the small band selected remains centered at 3600 cps. When the control signal rises to a value corresponding to 3800 cps or better, relay #2 closes and the small band centered at 4000 cps is selected. When any given relay is closed all lower numbered relays remain closed.

Let us consider the design of the filter. First assume that the Twin T is completely symmetric, i.e., X_1 X_2 X_o Y_1 Y_2 Y_o 1. The gain for the amplifier is found from Q $\frac{\text{Gain}}{4}$. Here there appears to be a contradic-

tion. As has been stated, the Q of this circuit must vary linearly with frequency, to maintain a constant bandpass. If the Twin T is completely symmetric at all center frequencies, then the Q of the filter is equal to $\frac{\text{Gain}}{4}$ and thus for a flat amplifier the Q would remain constant over the passband. It will be shown that by modification of the Twin T, the filter will have a Q that varies linearly with frequency.

Select the highest Q needed in the sub-band. For the #4 filter, the maximum $Q = \frac{f}{\Delta f} = \frac{5600 \text{ cps}}{300 \text{ cps}} = 18.67$. Then the required gain of the amplifier is $G = 4Q = 74.68$. The required gain is obtained from the amplifier stage with a 15K resistor in the plate circuit.

The values of the resistances and capacitors in the Twin T must now be chosen. Components must be picked subject to two constraints. The resistances of the Twin T must be of such a value that at any rejection frequency, the input resistance is large compared to the cathode follower resistance, and the output resistance smaller than one-third the size of the input impedance to the amplifier stage. The capacitor sizes should be large enough to swamp wiring capacitance and amplifier input capacitance. It must be stressed that the Twin T is very finely balanced and any change in the effective component values causes wide deviation from the desired operation.

The design equations for the Twin T components are:

$$R_1 = R_2, \; C_1 = C_2, \; C_3 = 2C_1 \qquad R_3 = \frac{R_1}{2}$$

$$f_{\text{rejection}} = \frac{1}{2\pi R_1 C_1}$$

For $f_r = 5600$ cps and $C_1 = 500 \mu\mu f$;

$$R_1 = 56.9K \qquad R_3 = 28.45K$$

Experimentally, it has been found that the Q of the filter with the values shown above being utilized is lower than the desired design value. To obtain the desired Q, reduce R_3 in the Twin T to approximately one-half of its design value. This changes the rejection frequency by a proportion which is of the same order as the Q. If R_1 and R_2 are now increased slightly, the rejection frequency returns to the desired value. The amplitude of the output must be of some desired level and minor modifications in R_1, R_2, and R_3 will allow this requirement to be met. For comparison, the designed and actual circuit values are shown below:

	Design	Actual
$R_1 = R_2 =$	56.9K	62.0K
$R_3 =$	28.45K	16K
$C_1 = C_2 =$	500 $\mu\mu f$	500 $\mu\mu f$
$C_3 =$	1000 $\mu\mu f$	1000 $\mu\mu f$

The filter is then tuned for the next rejection frequency, 5200 cps, by varying the components of the Twin T to obtain the desired rejection frequency, bandpass, and output level.

The remaining filters are of the same design with minor modifications. Triodes were used at low frequencies as the input capacitance was not as large a problem as it was in the upper sub-bands. Various short cuts were used in the other filters to ease the fine tuning requirements on the Twin T. The amplitude of the output may be quickly adjusted by varying the grid resistance of the amplifier stage. Insertion of a resistance in the feedback loop to the grid the cathode follower varies the Q of the system.

Resistance variation in the Twin T was chosen instead of capacitive variation for practical reasons. This resulted in more extensive tuning of the network as it will be noted from equation (4) that if the R's are varied to obtain the different rejection frequencies, the input impedance

to the Twin T varies with rejection frequency. Whereas, if capacitors are used as the variable, the input impedance to the Twin T remains constant.

The advent of Vericaps of the required size will remove the need for the relays and transistor relay control networks. The simplicity and performance of the resulting circuit will be excellent. The Twin T will contain six components instead of an entire matrix of elements. Capacitance variation will remove the necessity of juggling the system to obtain amplitude output equality as the Twin T input impedance will be invariant with rejection frequency. A Q linear with frequency may be obtained by having the series and parallel capacitors vary linearly with control voltage but at slightly different slopes.

In passing, an important comment must be made. In operation the action of the frequency information control signals is quantized. That is, the control signals are continuous in nature, but the passbands of the various filters shift in discrete steps only. The extent to which the intelligibility of the speech processing scheme is effected by this quantization must be determined by an additional investigation in which a continuous system, using voltage variable capacitors of the proper size, is developed and used for comparison. If the continuous system is not markedly more efficient than the quantized system now being investigated, then further bandwidth compression may be achieved by a quantization of the control signals.

Figure 33 is a photograph of the four voltage variable filter units and the modulator unit. Figure 34 is a photograph of the complete laboratory set-up for the speech processing system.

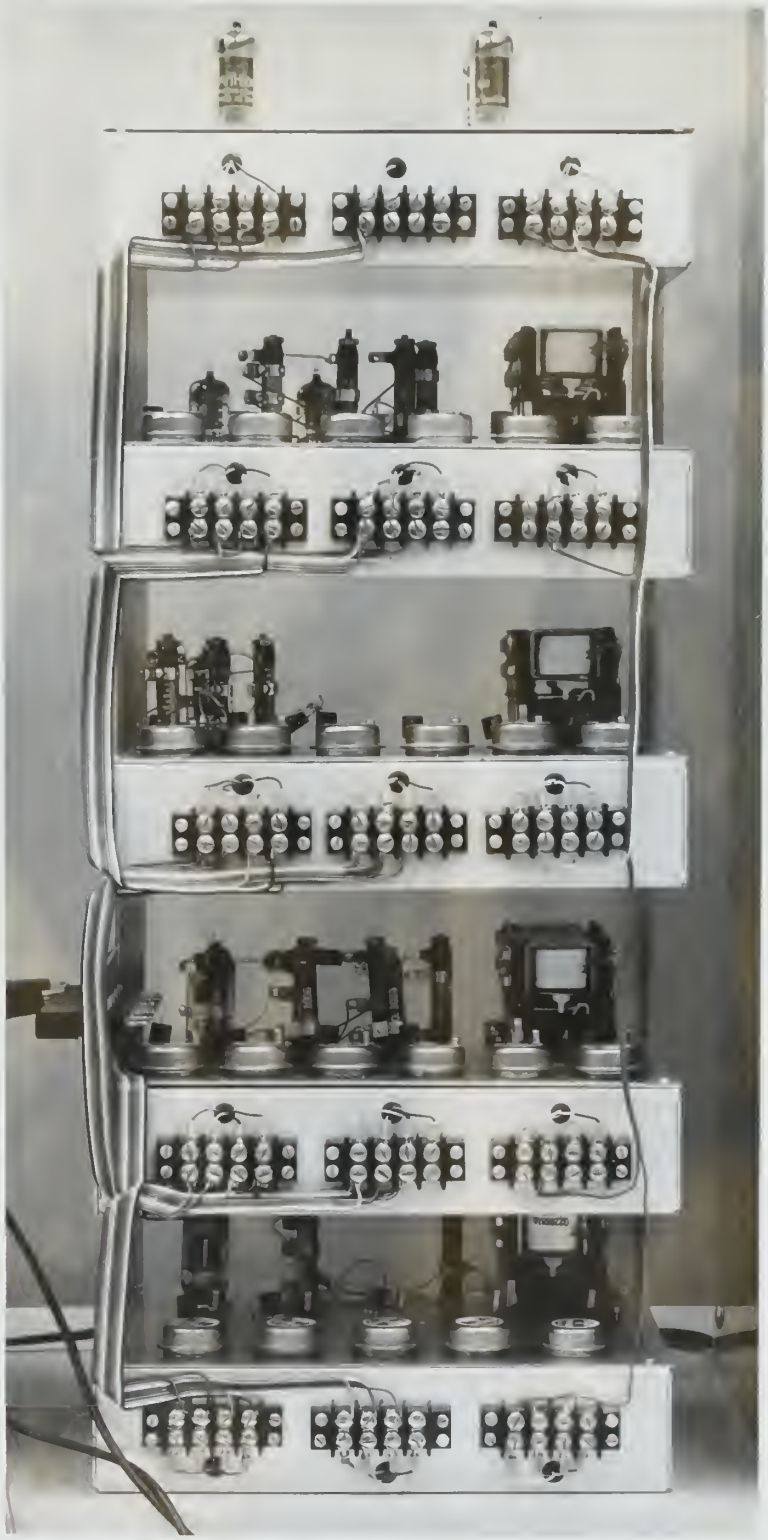


Figure 33. Voltage variable filter units and modulator unit. Top unit is modulator unit. Bottom four units are the four voltage variable filters.

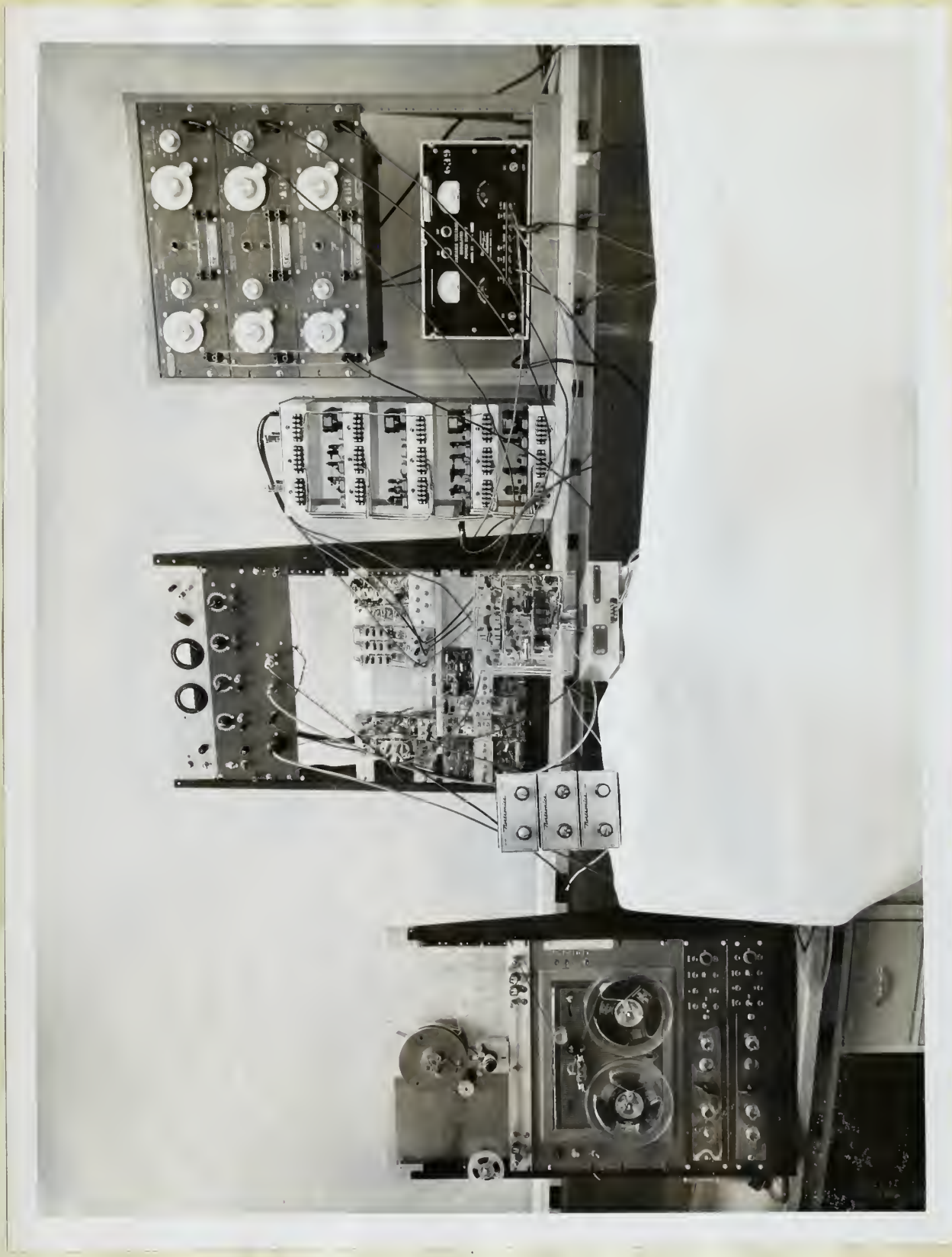


Figure 34. Laboratory set-up for speech processing system.

7. Conclusions and Recommendations.

In retrospect it must be stated that the investigation presented in this paper is but Phase One of a speech processing bandwidth compression scheme development and evaluation effort. Phase One consisted of the conceptual evolution of the system, a laboratory implementation, and a successful feasibility demonstration. Unfortunately, time limitations on the investigation were such that extensive quantitative results were not obtained. Qualitative results and the performance of the system were better than anticipated and were such that the feasibility of successfully exchanging voice information in a highly compressed bandwidth using the given system was definitely demonstrated.

In order to adequately describe the qualitative results obtained three things must first be discussed: First, the state of the system during the testing period; Second, the environment in which the testing was done; and Third, the development of a qualitative intelligibility scale for use in adequately discribing the results.

Trouble shooting of the system was far from complete when the system was tested. Severe mismatches between elements of the system were found to exist. Efforts to partially eliminate the mismatches resulted in vast improvements in the intelligibility of the system. The level of intelligibility achievable in a matched trouble-free system is still one of conjecture.

Testing was done in a very noisy environment. The clicking of the relays of the voltage variable filters forced conversationalists to raise their voices in the area of the system in order to be understood.

In order to most clearly describe the intelligibility of the system the following scale which describes given intelligibility levels associat-

ed with given physical environments.

Intelligibility Level	Physical Situation
A	Quiet room non-bandlimited speech, speaker recognition.
B	Noise room, non-bandlimited speech, speaker recognition.
C	Quiet room, bandlimited speech, speaker recognition, i.e. telephone communication.
D	Noisy room, bandlimited speech, speaker recognition.
E	Slight speech distortion, speaker recognition no effort to recognize words.
F	Slight distortion with noise, speaker recognition, only very slight effort to recognize words.
G	Distortion and noise such that speaker is not recognizable, very mild effort to recognize words.
H	Medium distortion and noise, speaker non-recognition, slight effort to recognize words.
I	Distortion and noise such that measurable effort is required for word recognition.
J	Distortion and noise such that severe effort is required for word recognition.
K	Distortion and noise such that many words are not recognized in connected text.
L	Very few words recognized and then only by extreme effort.
M	Total non-recognition.

Several listeners were utilized in testing the intelligibility of the system. The sound inputs to the system, which consisted of words, vowels, and other sounds, were recorded on magnetic tape and played into the system so that the listeners could only hear the output of the system. The listeners were given no clue as to what sounds to expect. Words in context were not used. The listeners were then asked to identify the synthesized sounds coming from the system. The evaluation of the system showed that for approximately 50% of the test words the intelligibility level corresponded to level "H" above. The remainder of the test words had a level of "I". Certain words were found to be extremely intelligible. Some of these were: six, international, avis, nine, and corporation. These words required no

effort for recognition. The vowel sounds were found to have a higher intelligible level, "G". The plosive sounds averaged between levels "G" and "H". This was better than anticipated. Inasmuch as the plosives have a rapid onset time it was thought that the smoothing action of the integrators, filters, etc., would reduce their intelligibility. The fact that they were better than anticipated is attributed to the discrete action of the voltage variable filter. It is believed that the transients set-up when entire units of resistors are switched in and out of the filter have onset characteristics similar to the plosive onset.

The RC time constants of the system were such that each of the seven control signals was limited to a maximum variation rate of 20 cps. The fastest rise time for the control signals was observed to be 20 milliseconds which corresponds to a low pass filter characteristic with a cutoff of 17.5 cps. For seven control signals each with a 20 cps bandwidth the total bandwidth for the system is 140 cps. This is a 25:1 reduction over the 3500 cps voice bandwidth commonly associated with SSB.

The goal of system silence between words was achieved and no speaker recognition was accomplished by the test listeners.

Further investigation of the sound generators of the synthesizer and the pitch synthesis technique is recommended. The sound generators should be a homogenious source of voice characterized bandlimited sound. Two techniques were utilized to develop a recorded source of this type of excitation. The first technique consisted of having one speaker talk through a bandlimited filter onto a continuous loop of magnetic tape while the loop cycles past the write head many, many times. It was found that the linear addition of sound on the tape hoped for was extremely difficult to achieve. The second technique consisted of having several speakers talk through a

bandlimited filter simultaneously and recording during only one cycle of the tape. This system was found to be far superior to the first technique. But, much investigation is still required to determine the optimum means for implementing the sound generator concept.

It is recommended that an investigation of the possibility of using a pitch oscillator whose frequency is controlled by the Pitch Control Signal to synthesize the pitch frequency be conducted. The pitch oscillator would replace the 100 to 200 cps sound generator and the voltage variable filter associated with the pitch channel. A system test utilizing the pitch oscillator will determine if the intelligibility of the system is enhanced.

System recommendations, aside from the obvious one of system matching the various elements of the system, are optimization of:

1. Channel frequency limits placement.
2. Bandwidth of voltage variable filters.
3. Relative amplitude levels of the sound generators.

Investigation is still required to determine if the channels selected by the analyzer filter bank are optimum with respect to frequency limits and bandwidth. Perhaps the lowest channel should not be from 300 to 1500 cps but should be from 200 to 1000 cps. The proper channel width and frequency limits can only be optimized by further intensive investigation. Also further investigation should be done on the possibility of extracting amplitude and frequency information from different areas in the frequency spectrum.

Optimization of the width of the bandpass of the voltage variable filters is required. Testing of the system should be done using different bandwidths to determine the best bandwidth to use.

During the testing of the system it was found that better intelligi-

bility resulted if the amplitude levels of the sound generators were not the same. Further research is required to determine the optimum relative amplitude levels of the sound generators.

The speech processing system provides an excellent level of transmission security in itself. An enemy cannot reconstruct speech from the transmitted control signals unless he knows the exact function of each of the seven control signals and can duplicate the system synthesizer. Further security can be achieved by multiplexing techniques and by time and frequency scrambling of the control signals.

BIBLIOGRAPHY

1. Greefkes, J. A., and F. de Jager, "'Frena', A System of Speech Transmission at High Noise Levels," Philips Technical Review, 19, No. 3, pp 73-83, 1957/58.

2. Dudley, H., "The Carrier Nature of Speech," B.S.T.J., 19, p. 495, Oct. 1940.

3. Stetson, R. H., Motor Phonetics, North-Holland Pub. Co., Amsterdam, 2nd Ed., 1951.

4. Cherry, C., On Human Communication, The Technology Press of M.I.T., 1957.

5. Kaiser, L., Manual of Phonetics, North-Holland Pub. Co., Amsterdam, 1957.

6. Campanella, S. J., "A Survey of Speech Bandwidth Compression Techniques," IRE trans on audio, AU-6:5, pp. 104-116, Sept.-Oct. 1958.

7. Schouten, J. F., "The Perception of Pitch, " Philips Technical Review, 5:10, pp. 286-294, 1940.

8. Steinberg, J. C., "Effects of Distortion upon Recognition of Speech Sounds," J. Acoust. Soc. Amer., 1, pp. 121-137, 1929.

9. Stevens, S. S., and J. Volkmann, "The Relation of Pitch to Frequency," Am. J. Psychol., 53, pp. 329-353, 1940.

10. Rayleigh, Lord, Theory of Sound, Macmillan and Co., Ltd., London, 1896.

11. Licklider, J. C. R., "The Intelligibility of Amplitude-Dichotomized, Time-Quantized Speech Waves," J. Acoust. Soc. Amer., 22, pp. 820-823, November 1950.

12. University College (London), Communication Research Centre, Studies in Communication, Martin Secker and Warburg, Ltd., London, 1955.

13. Fairbanks, G., W. L. Everitt, and J. P. Laeger, "Method for Time or Frequency Compression-Expansion of Speech," 1953 IRE Conv. Record, pt. 8, pp. 120-124.

14. David, E. E., and H. S. McDonald, "Note on Pitch Synchronous Processing of Speech," J. Acoust. Soc. Amer., 28, pp. 1201-1266, November 1956.

15. Dudley, H., "Remaking Speech," J. Acoust. Soc. Amer., 11, pp. 169-177, October 1939.

16. Dunn, H. K., and H. L. Barney, "Artificial Speech in Phonetics and Communication," Bell Telephone System Monograph No. 3020, July 1958.

17. Howard, C. R., "Speech Analysis-Synthesis Scheme Using Continuous Parameters," J. Acoust. Soc. Amer., 28, pp. 1091-1098, November 1956.

18. Reitz, R. R., "Apparatus for Finding the Pitch Frequency in a Complex Wave," Patent No. 2,593,698, April 22, 1952.

19. Peterson, E., "Wave Analyzer for Determining the Fundamental Frequency of a Complex Wave," Patent No. 2,593,694., April 22, 1952.

20. Miller, R. L., "Determination of Pitch Frequency of Complex Wave," Patent No. 2,627,541., February 3, 1953.

21. Tarnoczy, T. H., "Determination of the Speech Spectrum through Measurements of Superposed Samples," J. Acoust. Soc. Amer., 28, pp. 1270-1275, November 1956.

22. Dolansky, L. O., "Electronically Controlled Audio Filters," IRE Conv. Record, pt. 7, pp. 41-48, 1955.

23. Fant, C. G. M., "A Continuously Variable Filter," J. Acoust. Soc. Amer., 22, pp. 449-453, July 1950.

24. Linvill, J. C., "Transistor Negative Impedance Converters," Proc. IRE 41, pp. 725-729, June 1953.

25. Wolf, A., "Note on Parallel T Resistance-Capacitance Network," Proc. IRE, 34.9, p 629, September 1946.

26. Tucker, M. J., and L. Draper, "A High Q RC Feedback Filter for Low Audio Frequencies," Electronic Engineering, 27, p. 451, October 1955.

27. Scott, H. H., "A New Type of Selective Circuit and Some Applications," Proc. IRE, 26.2, p. 226, February 1938.